Rev. 10/93

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

IN THE APPLICATION OF:

SAVERIO CARL FALCO ET. AL.

CASE NO.: BB1037USCNT APPLICATION NO.: 10/804678

CONFIRMATION NO.: 9737 GROUP ART UNIT: 1638

EXAMINER: E. F. MCELWAIN FILED: MARCH 19, 2004

FOR: LYSINE-INSENSITIVE ASPARTOKINASE GENE AND METHOD FOR INCREASING THE LYSINE AND THREONINE CONTENT OF THE SEEDS OF PLANTS

Brief on Appeal

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is an appeal of the Final Rejection, mailed September 28, 2007, of Claims 39-53 of the above-identified application.

TABLE OF CONTENTS

(I) Real Party in Interest

The real party in interest in this Appeal is E. I. du Pont de Nemours and Company, the assignee of the entire right, title and interest of the above-identified patent application.

(II) Related Appeals and Interferences

There are no related Appeals or Interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the pending Appeal.

(III) Status of Claims

Claim 1-20 were originally filed.

Claims 21-38 were added during prosecution and then claims 1-38 were cancelled. Claims 39-53 were added during prosecution and were rejected. There are three independent claims: 39, 44 and 49.

The currently pending and appealed claims are claims 39-53 which are set forth in the Claims Appendix attached hereto.

(IV) Status of Amendments Filed Subsequent to Final Rejection

A Response After Final was filed electronically on February 4, 2008. Claim 44 was the only claim amended. The Response After Final was entered as set forth in the Advisory Action dated March 5, 2008.

(V) Summary of the Invention

Lysine-ketoglutarate reductase (LKR) and saccharopine dehydrogenase (SDH) catalyze the first and second steps, respectively, in the breakdown pathway of lysine resulting in the product of saccharopine or alpha-amino adipic acid. Thus, the ability to down-regulate expression of the LKR/SDH gene can lead to an increase in the level of lysine in seed by preventing, either partially or fully, the breakdown of lysine.

Claim 39 of the instant invention relates to a chimeric gene capable of causing an increased level of lysine in seeds obtained from a transformed plant, the chimeric gene comprising:

a) an isolated nucleic acid fragment comprising a nucleic acid sequence which is useful in antisense inhibition or sense suppression of endogenous lysine ketoglutarate reductase/saccharopine dehydrogenase activity in a plant or plant cell wherein said isolated nucleic acid fragment comprises all or a part of the nucleic acid sequence encoding a plant lysine ketoglutarate reductase/saccharopine dehydrogenase, said part being sufficient in length for use in antisense inhibition or sense suppression; and

b) at least one regulatory sequence operably linked to said fragment.

Also claimed are plants transformed with this chimeric gene, seeds obtained from such transformed plants and a method for increasing the lysine content in a plant seed using this chimeric gene.

This is discussed in the specification on page 31 at line 30 through the last line on page 37 and in Example 20 on pages 92-98.

Claim 44 is virtually identical to claim 39 with the exception that the transformed plant is a corn plant.

This is discussed in the specification on page 31 at line 30 through the last line on page 37 and in Example 20 on pages 92-98.

Claim 49 is similar to claim 44 in that the transformed plant is a corn plant. Claim 49 recites that the isolated nucleic acid fragment comprises all of a part of the nucleic acid fragment of SEQ ID NO:120 which is the sequence of a 3,265 nucleotide cDNA from corn.

This is discussed in the specification on page 8, line 24, page 31 at line 30 through the last line on page 37 and in Example 20 on pages 92-98.

(VI) Grounds of Rejection To Be Reviewed on Appeal

There are two grounds of rejection presented for review:

(a) Whether claims 39-53 comply with the written description requirement under 35 USC §112, first paragraph, in view of the following:

(i) diagrams (sequence alignments) that were not part of the specification but contain sequence(s) that are disclosed in the specification;

(ii) two post-filing date publications (in view of the priority claimed) that were not available at the time of the invention (one of which was co-authored by the above-identified co-inventors) but provide information about the sequences disclosed and claimed in the instant application; and

(iii) two Declarations of Dr. Carl Falco, one of the co-inventors of the subject application purportedly because only one example of a sequence that functions is provided since "one example of a sequence is not sufficient to support the claimed genus of any nucleic acid sequence which is useful in inhibition of LKR/SDH activity in a plant or plant cell...."

(b) Whether claims 39-53 comply with the enablement requirement under 35 USC §112, first paragraph, in view of the two Declarations of Dr. Carl Falco and sequence alignment, sequence alignments and two post-filing date publications, one of which was co-authored by Dr. Falco and Dr. Epelbaum, the co-inventors of the subject application.

(VII) Argument

(a) The rejection of claims 39-53 under 35 USC §112, first paragraph, as failing to comply with the written description requirement.

Drs. Falco and Epelbaum, the co-inventors of the claimed invention, were the first to report the molecular cloning of a plant LKR/SDH genomic and cDNA sequence. They subsequently co-authored a paper (Epelbaum et al., Plant Molecular Biology 35:735-748 (1997)) that was published subsequent to the filing

of the above-identified application. A copy of this paper was previously submitted and is attached hereto as Evidence Appendix A.

Epelbaum et al. and Example 20 on page 94, describes the isolation of the gene encoding LKR/SDH from an *Arabidopsis thaliana* genomic DNA library based on the homology between the yeast biosynthetic genes encoding SDH (lysine-forming) or SDH (glutamate- forming) and Arabidopsis expressed sequence tags.

Primers were designed from these expressed sequence tags (ESTs) (page 736 of the paper under "Materials and Methods", section "Gene Isolation"). The sequences of these ESTs, T13618 and T45802, correspond to SEQ ID NOs: 102 and 103, respectively, of the instant specification (page 32, 3rd paragraph and example 20, page 94, last paragraph).

The sequences of ESTs served as the basis for designing primers (SEQ ID NOs:108 and 109) for use in the PCR amplification of a 2.24kb DNA fragment form genomic Arabidopsis DNA (specification, page 32 , third paragraph, and page 95, first paragraph, and Epelbaum et al., page 736 "Materials and Methods" section named "Gene isolation).

The 2.24 kb DNA fragment was then used to isolate a larger genomic DNA fragment. The sequence of this larger genomic fragment is provided in SEQ ID NO:110 of the specification and corresponds to the nucleotide sequences shown in Figure 2 of the Epelbaum et al. paper. Subsequently the full length DNA coding sequence for the Arabidopsis LKR/SDH was isolated via RT-PCR. The sequence of the Arabidopsis LKR/SDH cDNA is provided in SEQ ID NO:111 of the instant specification and is indicated by capital letters in the nucleotide sequence in Fig.2 of the paper. The deduced amino acid sequence of the Arabidopsis LKR/SDH protein is shown in SEQ ID NO:112 and corresponds to the amino acid sequence shown in Fig.2 of the Epelbaum et al. paper.

The deduced amino acid sequence set forth in Figure 2 on pages 739-741 of Epelbaum et al. shows that in plants the LKR/SDH activities are carried on a single bi-functional protein. Function of the Arabidopsis LKR/SDH protein can be

assayed using previously described assays with some minor modifications (Page 738 of paper under "LKR specific activity", "SDH specific activity"). The "SDH portion" (SEQ ID NO:131 of the instant specification) of the bi-functional Arabidospis LKR/SDH protein could be successfully expressed and assayed in E.coli.

Accordingly, what is discussed in the Epelbaum et al. paper relates directly to the sequence and subject matter of the instant application. Even though this paper was published after the priority of the instant application, it simply further discusses the sequence already disclosed and claimed in the instant application.

Figure 4 on page 744 of the Epelbaum paper sets forth a comparison of the deduced amino acid sequences of three fungal genes encoding SDH (lysine forming) with the *A. thaliana* LKR.

Figure 5 is a comparison of the deduced amino acid sequence of the *S. cerevisiae* SDH (glutamate forming) and the *A. thaliana* SDH. The *Arabidopsis* sequences used in these comparisons are the same Arabidopsis sequences disclosed in the instant application. In fact, the comparison in Figure 5 of Epelbaum is similar to the comparison in Figure 9 of the instant application. Figure 9 is described on page 10 at lines 1-2 of the instant specification as showing "the amino acid similarity between the polypeptides encoded by two plant cDNAs and fungal *S. cerevisiae* (glutamate forming)." Figure 9 is also discussed in Example 20 on page 95 at lines 1-3 of the instant application.

Based on comparison of the Arabidopsis LKR and SDH with other LKR and SDH proteins, as mentioned above, degenerate primers (SEQ ID NOS:113 and 114) were designed and additional LKR and SDH sequences from corn and soy were identified and isolated (page 95, last paragraph through page 96 first paragraph of instant application). Subsequently, near full length soy and corn LKR/SDH sequences were obtained. The comparison of the corn and soy LKR/SDH sequences with ESTs from other plants enabled the identification and

isolation of sequences from rice and wheat (described on page 95 at line 33 through the end of page 96).

A cosuppression experiment using a modified shorter version (1268bp fragment, see below) of the corn LKR/SDH (SEQ ID NO:120), is discussed in the specification starting on page 97 at lines 15-36.

Dr. Falco's declaration(s) provided additional data showing that the1268 bp gene fragment include the LKR coding domain obtained from the corn LKR-SDH sequence (SEQ ID NO:120) was successfully used in cosuppression studies to produce seeds having increased accumulation of lysine. This increase in lysine appeared to be directly related to the co-suppression of LKR/SDH.

Dr. Falco's Declaration dated August 24, 2000 (copy provided in Evidence Appendix B) shows that two important elements that are necessary and sufficient to practice the invention are provided: (1) the motivation to "knock out" LKR (as is set forth in paragraph 4 of Dr. Falco's declaration[,]) and (2) disclosure of the first nucleic acid fragments encoding a plant LKR. With these fragments in hand, then it was possible to isolate LKR fragments from any other plant desired, and use them to block expression utilizing antisense inhibition and/or cosuppression. Dr. Falco's declaration demonstrates that blocking the first step in lysine catabolism, i.e., "knocking out" LKR, leads to increased accumulation of lysine in seeds.

Dr. Falco's Declaration dated February 16, 2001, (copy provided in Evidence Appendix C) one of the co-inventors of the subject case, sets forth data showing seeds with increased lysine content that were obtained from plants co-transformed with DHDPS and LKR. The LKR sequence, a 1268 bp gene fragment of obtained from the sequence comprising the near full length corn LKR/SDH (SEQ ID NO:120), was successfully used to increase lysine and correlated with cosuppression of LKR/SDH.

The experiments discussed in Dr. Falco's previously submitted declaration taken together with the detailed description of the invention provided in the patent application and the previous declaration (dated August 24, 2000), clearly

demonstrate that an increased lysine content is achieved when a foreign lysine insensitive DHDPS gene (with or without a lysine insensitive AK gene) is combined with a foreign co-suppressing LKR gene.

Another reference that demonstrates that the nucleotide sequences described in the invention encode plant lysine-ketoglutarate reductase and saccharopine dehydrogenase proteins are Tang et al., Plant Cell **9**:1305-1316 (1997) entitled "Regulation of lysine catabolism through lysine-ketoglutarate reductase and saccharopine dehydrogenase in Arabidopsis" (copy provided in Evidence Appendix D).

This paper reports the cloning of an Arabidopsis cDNA encoding a bifunctional polypeptide that contains both of these enzymatic activities linked to each other.

The Arabidopsis sequence dislosed by Tang et al. is essentially identical to SEQ ID NO:111 of the instant application. Tang et al. page 1308, right-hand column, discloses that bacterial cells transformed with a plasmid having the LKR and SDH insert showed SDH, but no LKR activity. However, yeast cells transformed with a plasmid having the LKR insert had significantly higher LKR activity than control yeast cells transformed with the same plasmid lacking the LKR insert.

Given that the Arabidopsis sequence disclosed in the instant application, SEQ ID NO:111, is essentially identical to that disclosed by Tang et al., then it would be expected that SEQ ID NO:111 would also produce LKR activity if expressed in yeast as described by Tang et al.

Structural and functional properties of the bifunctional LKR/SDH enzyme are discussed in the Tang et al. paper, starting on page 1312, left hand column. This is the same enzyme disclosed in the instant application.

Analysis of LKR and SDH activities is described on page 1315, left hand column, and it should be clear to those skilled in the art that such analysis is clearly within the skill in the art.

It is respectfully submitted that in view of Epelbaum et al. and Tang et al., it should be clear that there is a correlation between sequence similarity and functionality insofar as LKR/SDH activity is concerned.

Furthermore, it should be reiterated that the information regarding the full length *Arabidopsis* LKR/SDH nucleotide and amino acid sequences, the expression and assay of the SDH portion of the *Arabidopsis* LKR/SDH, the soy and corn partial LKR/SDH sequences and the use of LKR for cosuppression experiments was available as of March 27, 1997 in priority application having application no.: 08/824,6127.

Accordingly, it is respectfully submitted that the claims fully comply with the written description requirement of 35 USC §112, first paragraph.

(b) The rejection of claims 39-53 under 35 USC §112, first paragraph, as failing to comply with the enablement requirement.

It is believed that all of the foregoing discussion, references and information discussed above with respect to written description rejection, are equally apposite with respect to the enablement rejection of claims 39-53 under 35 USC §112, first paragraph.

Specifically, it is stated on page 4 of the Office Action mailed on January 25, 2007 that the "specification does not demonstrate that any of the claimed sequences have homology to saccharopine dehydrogenase (SDH) and that SEQ ID NO: 120 and 122 are not full length sequences (page 34). Therefore, it is even more uncertain that the claimed sequences would encode the portions required to confer LKR activity. "

It was further stated on page 5 of this same Office Action that "De Luca teaches that modifying plant biosynthetic pathways by transforming plants with genes encoding enzymes involved in a biosynthetic pathway is highly unpredictable and often the desirable results are impossible to achieve".

It is respectfully submitted that ample information is available in case of the lysine biosynthetic and catabolic pathways that clearly demonstrates how to increase lysine production via modification of the biosynthetic and catabolic pathways. The use of lysine feedback-insensitive versions of the key biosynthetic enzymes, DHDPS and AK, has been shown to lead to an increase in free lysine levels. The instant specification teaches that blocking the first step in lysine catabolism will lead to increased accumulation of lysine.

This taken together with all of the information discussed above with respect to the written description rejection, the it is respectfully submitted that one of ordinary skill in the art would be able to practice the claimed invention without engaging in undue experimentation.

As was discussed above, Dr. Falco's Declaration dated February 16, 2001, (copy provided in Appendix C) one of the co-inventors of the subject case, sets forth data showing seeds with increased lysine content that were obtained from plants co-transformed with DHDPS and LKR. The LKR sequence, a 1268 bp gene fragment of obtained from the sequence comprising the near full length corn LKR/SDH (SEQ ID NO:120), was successfully used to increase lysine and correlated with cosuppression of LKR/SDH.

The experiments discussed in Dr. Falco's previously submitted declaration taken together with the detailed description of the invention provided in the patent application and the previous declaration (dated August 24, 2000), clearly demonstrate that an increased lysine content is achieved when a foreign lysine insensitive DHDPS gene (with or without a lysine insensitive AK gene) is combined with a foreign co-suppressing LKR gene.

Reference was made to Doerks (TIG14, no. 6:248-250, June 1998) (copy provided in Evidence Appendix E) for the proposition that sequence homology is not sufficient to predict function of an encoded sequence.; reference was made to Smith et al. (Nature Biotechnology 15:P1222-1223, November 1997) (copy provided in Evidence Appendix F) for the proposition that homologuos proteins can have different functionality; reference was made to Brenner (TIG 15, 4:132-

133, April 1999) (copy provided in Evidence Appendix G) which discusses the problem of inferring function from homology and Borks (TIG12, 10:425-427, Ovotber 1996) (copy provided in Evidence Appendix H) which teaches problems with sequence databases that can result in the misinterpretation of sequence data.

Given what has been discussed herein, it is respectfully submitted that there is no such problem with respect to the claimed invention.

Attached hereto is Evidence Appendix I which is an alignment of the LKR domains of the plant bifunctional LKR/SDH proteins from Arabidopsis (SEQ ID NO:112), corn (SEQ ID NO:122, encoded by SEQ ID NO:120) and soybean (SEQ ID NO:121) and the monofunctional lysine-forming SDH proteins from *S.cerevisiae* (gi:453184), *C.albicans* (gi:1170847) and *Y.lipolytica* (gi:173262).

Evidence Appendix J (submitted herewith) is comparison of the SDH domains of the bifunctional plant LKR/SDH proteins from Arabidopsis (SEQ ID NO:112), corn (SEQ ID NO:122, encoded by SEQ ID NO:120) and soybean (SEQ ID NO:121) and the monofunctional glutamate-forming SDH protein from *S.cerevisiae* (gi:729968). Residues that are identical among at least one of the plant sequences and at least one of the yeast sequences are indicated by an asterisk above each alignment. Residues that are identical among at least two plant sequences are indicated by a plus sign above each alignment.

The plant LKR domains share about 70% and 60% sequence identity with each other, respectively, whereas the plant LKR domains and yeast lysine-forming SDH proteins share between 15% and 17% sequence identity.

The plant SDH domains share about 60% sequence identity among each other and around 30% sequence identity with the yeast protein. Alignments and percent identity calculations were performed using the Clustal V method of alignment.

The comparisons set forth in Evidence Appendices I and J demonstrate that the sequences of the invention possess stretches of highly conserved regions. One skilled in the art would appreciate that the more highly conserved a

residue is, the less likely that it could be modified and function maintained. From these alignments, one could quickly determine which amino acid residues might be modified in SEQ ID NO:122 (encoded by SEQ ID NO:120) without a likely change in function.

In the instant specification, the cDNA fragments of the bifunctional Arabidopsis LKR/SDH were identified based on the homology to the monofunctional proteins from yeast. The sequence similarity between the yeast and plant polypeptides (Fig.9 of instant specification) demonstrated that these cDNAs encode Arabidopsis saccharopine dehydrogenase.

The complete genomic sequence of the Arabidopsis LKR/SDH gene was subsequently isolated and the cDNA sequence and corresponding amino acid sequence determined. The LKR/SDH cDNA revealed an ORF of 3.16 kb, which predicts a protein of 117 kd, and confirms that the LKR and SDH enzymes reside on one polypeptide.

In order to isolate further plant LKR/SDH sequences, degenerate primers based upon comparison of the Arabidopsis LKR/SDH amino acid sequence with that of other LKR proteins were designed. These were used to amplify soybean and corn LKR/SDH fragments using PCR from mRNA, or cDNA synthesized from mRNA, isolated from developing soybean or corn seeds. Near full length sequences for the LKR/SDH sequences were obtained using Race and hybridization protocols. Furthermore, partial rice and wheat were isolated based on homology to the Arabidopsis protein.

Evidence Appendix K (submitted herewith) is an alignment of the plant bifunctional LKR/SDH proteins from Arabidopsis (SEQ ID NO:112), corn (SEQ ID NO:122) and soybean (SEQ ID NO:121). Amino acid residues identical among at least two plant sequences are indicated by an asterisk on the top row; dashes are used by the program to maximize the alignment of the sequences. The LKR and SDH domains have been boxed in Evidence Appendix K to facilitate review of the enclosed Evidence Appendix K. It should also be noted that, in addition to

the LKR and SDH domains, a high degree of homology is also observed in the intermediary or 'spacer' region of the bifunctional LKR-SDH polypeptide.

(VII) Conclusion

When this is viewed in combination with the information presented in Epelbaum et al. (discussed above), one is inexorably led to the conclusion that one skilled in the art can make and use the claimed invention without engaging in undue experimentation.

Accordingly, the Board is respectfully requested to reverse the final rejection of pending claims 39-53 and indicate allowability of all claims.

Enclosed herewith is a Petition for a three (3) month extension of time to permit the filing of the Brief on Appeal. Please charge the fee for the extension of time of three (3) months, as well as the requisite fee set forth in 37 CFR §1.17(f), to Appellant's Assignee's (E. I. du Pont de Nemours and Company) Deposit Account No. 04-1928.

Respectfully submitted,

/Lynne M. Christenbury/

LYNNE M. CHRISTENBURY
ATTORNEY FOR APPLICANTS
Registration No.: 30,971
Telephone: (302) 992-5481
Facsimile: (302) 892-1026

Dated: June 13, 2008

Claims Appendix

Claim 39. (previously presented) A chimeric gene capable of causing an increased level of lysine in seeds obtained from a transformed plant, the chimeric gene comprising:

a) an isolated nucleic acid fragment comprising a nucleic acid sequence which is useful in antisense inhibition or sense suppression of endogenous lysine ketoglutarate reductase/saccharopine dehydrogenase activity in a plant or plant cell wherein said isolated nucleic acid fragment comprises all or a part of the nucleic acid sequence encoding a plant lysine ketoglutarate reductase/saccharopine dehydrogenase, said part being sufficient in length for use in antisense inhibition or sense suppression; and

b) at least one regulatory sequence operably linked to said fragment.

Claim 40. (previously presented) A plant comprising the chimeric gene of Claim 39 in its genome.

Claim 41. (previously presented) Seed obtained from the plant of Claim 40.

Claim 42. (previously presented) A method for increasing lysine content in a plant seed which comprises:

(a) transforming plant cells with the chimeric gene of Claim 39;

(b) regenerating fertile mature plants from the transformed plant cells obtained from step (a) under conditions suitable to obtain seeds;

(c) screening progeny seed of step (b) for increased lysine content; and

(d) selecting those lines whose seeds have increased lysine content.

Claim 43. (previously presented) Seed obtained by the method of Claim 42.

Claim 44. (previously presented) A chimeric gene capable of causing an increased level of lysine in seeds obtained from a transformed corn plant, the chimeric gene comprising:

a) an isolated nucleic acid fragment comprising a nucleic acid sequence which is useful in antisense inhibition or sense suppression of endogenous lysine ketoglutarate reductase/saccharopine dehydrogenase activity in a corn plant or corn plant cell wherein said isolated nucleic acid fragment comprises all or a part of the nucleic acid sequence encoding a corn plant lysine ketoglutarate reductase/ saccharopine dehydrogenase, said part being sufficient in length for use in antisense inhibition or sense suppression; and

b) at least one regulatory sequence operably linked to said fragment.

Claim 45. (previously presented) A corn plant comprising the chimeric gene of Claim 44 in its genome.

Claim 46. (previously presented) Seed obtained from the corn plant of Claim 45.

Claim 47. (previously presented) A method for increasing lysine content in a corn plant seed which comprises:

(a) transforming corn plant cells with the chimeric gene of Claim 44;

(b) regenerating fertile mature plants from the transformed corn plant cells obtained from step (a) under conditions suitable to obtain seeds;

(c) screening progeny seed of step (b) for increased lysine content; and

(d) selecting those lines whose seeds have increased lysine content.

Claim 48. (previously presented) Seed obtained by the method of Claim 47.

Claim 49. (previously presented) A chimeric gene capable of causing an increased level of lysine in seeds obtained from a transformed corn plant, the chimeric gene comprising:

a) an isolated nucleic acid fragment comprising a nucleic acid sequence which is useful in antisense inhibition or sense suppression of endogenous lysine ketoglutarate reductase/saccharopine dehydrogenase activity in a corn plant or plant cell wherein said isolated nucleic acid fragment comprises all or a part of the nucleic acid sequence of SEQ ID NO:120, said part being sufficient in length for use in antisense inhibition or sense suppression; and

b) at least one regulatory sequence operably linked to said fragment.

Claim 50. (previously presented) A plant comprising the chimeric gene of Claim 49 in its genome.

Claim 51. (previously presented) Seed obtained from the plant of Claim 50.

Claim 52. (previously presented) A method for increasing lysine content in a plant seed which comprises:

(a) transforming plant cells with the chimeric gene of Claim 49;

(b) regenerating fertile mature plants from the transformed corn plant cells obtained from step (a) under conditions suitable to obtain seeds;

(c) screening progeny seed of step (b) for increased lysine content; and

(d) selecting those lines whose seeds have increased lysine content.

Claim 53. (previously presented) Seed obtained by the method of Claim 52.

Evidence Appendix A

Epelbaum et al., Plant Molecular Biology **35**:735-748, 1997

This reference was entered into the record by the Examiner in the office Action mailed September 28, 2007, initialed PTO form 1449.

EVIDENCE APPENDIX A

Plant Molecular Biology 35: 735–748, 1997.

Lysine-ketoglutarate reductase and saccharopine dehydrogenase from *Arabidopsis thaliana*: nucleotide sequence and characterization

Sabine Epelbaum*, Raymond McDevitt and S. Carl Falco
*Agricultural Products, E.I. DuPont de Nemours & Co., Wilmington, DE 19880-0402, USA (*author for correspondence)*

Received 4 April 1997; accepted in revised form 7 July 1997

Key words: Arabidopsis thaliana, heterologous expression, lysine catabolism, lysine-ketoglutarate reductase saccharopine dehydrogenase

Abstract

We isolated the gene encoding lysine-ketoglutarate reductase (LKR, EC 1.5.1.8) and saccharopine dehydrogenase (SDH, ED 1.5.1.9) from an *Arabidopsis thaliana* genomic DNA library based on the homology between the yeast biosynthetic genes encoding SDH (lysine-forming) or SDH (glutamate-forming) and *Arabidopsis* expressed sequence tags. A corresponding cDNA was isolated from total *Arabidopsis* RNA using RT-PCR and 5′ and 3′ Race. DNA sequencing revealed that the gene encodes a bifunctional protein with an amino domain homologous to SDH (lysine-forming), thus corresponding to LKR, and a carboxy domain homologous to SDH (glutamate-forming). Sequence comparison between the plant gene product and the yeast lysine-forming and glutamate-forming SDHs showed 25% and 37% sequence identity, respectively. No intracellular targeting sequence was found at the N-terminal or C-terminal of the protein. The gene is interrupted by 24 introns ranging in size from 68 to 352 bp and is present in *Arabidopsis* in a single copy. 5′ sequence analysis revealed several conserved promoter sequence motifs, but did not reveal sequence homologies to either an Opaque 2 binding site or a Sph box. The 3′-flanking region does not contain a polyadenylation signal resembling the consensus sequence AATAAA. The plant SDH was expressed in *Escherichia coli* and exhibited similar biochemical characteristics to those reported for the purified enzyme from maize. This is the first report of the molecular cloning of a plant *LKR-SDH* genomic and cDNA sequence.

Abbreviations: AK, aspartate kinase; CTP, chloroplast transit peptide; DHDPS, dihydrodipicolinate synthase; LKR, lysine-ketoglutarate reductase; SDH, saccharopine dehydrogenase.

Introduction

Lysine is synthesized in higher plants and in many bacterial species from aspartate [6, 8]. Its rate of synthesis in plants is regulated mainly by feedback inhibition of aspartate kinase (AK) and dihydrodipicolinate synthase (DHPS) [6]. These enzymes therefore play an important role in determining the level of free lysine. Control of the biosynthetic pathway to lysine is of special interest, since lysine levels are low in the seeds of important crop plants, such as corn, thereby decreasing its nutritional quality [9].

Expression of feedback insensitive bacterial DHPS has been shown to result in elevated levels of free lysine in canola, soybean, and maize seeds [9, Falco *et al.*, unpublished results]. In each case the increased level of free lysine is accompanied by accumulation of the lysine breakdown products saccharopine or α-amino adipic acid. Lysine-ketoglutarate reductase (LKR, EC 1.5.1.8) and saccharopine dehydrogenase (SDH, EC 1.5.1.9) catalyze the first and second step, respectively, in the breakdown pathway of lysine that produces these intermediates in seeds (Figure 1) [1]. LKR condenses lysine and α-ketoglutarate into saccharopine

The nucleotide sequence data reported will appear in the GenBank, EMBL and DDBJ Nucleotide Sequence Databases under the accession numbers U95758 (*A. thaliana* (Landsberg *erecta*) *LKR-SDH* gene) and U95759 (*A. thaliana* (Columbia) *LKR-SDH* gene).

and SDH converts saccharopine to α-amino adipic-δ-semialdehyde. Biochemical and genetic evidence derived from human and bovine studies demonstrate that mammalian LKR and SDH enzyme activities are present on a single protein with a monomer molecular mass of 115 kDa [25]. Recent results obtained by Goncalves-Butruille *et al.* suggest that both enzyme activities from maize also reside on a single protein [15]. This contrasts with the fungal enzyme activities which are carried on separate proteins, SDH (lysine-forming) with a molecular mass of about 44 kDa and SDH (glutamate-forming) with a molecular weight of about 51 kDa [10, 12, 28, 34]. In fungi these enzymes catalyze the final two steps in the lysine biosynthetic pathway rather than a lysine catabolic pathway. Several genes for the fungal SDH's have been isolated and sequenced, but no plant or animal genes have yet been reported. There is little information on the regulation of lysine catabolism in plants. Evidence from studies on tobacco and maize suggest that LKR expression is developmentally regulated and in tobacco seeds LKR activity is stimulated through an intracellular signaling cascade involving calcium and protein phosphorylation, but the exact control mechanisms remain to be determined [1, 18, 19]. Nothing is known about the intracellular location of the lysine breakdown pathway. Lysine biosynthesis appears to be confined to the chloroplast [5, 26, 27].

In order to achieve a better understanding of the physiological role of lysine catabolism in higher plants, we have isolated and characterized the gene encoding LKR and SDH from the model plant *Arabidopsis thaliana*.

Materials and methods

Strains

The *E. coli* strains used were LE392, DH5α and BL219(DE3)pLysS [Novagen], the *Arabidopsis thaliana* ecotype Landsberg *erecta* and Columbia.

Gene isolation

Primers were designed from *Arabidopsis* expressed sequence tags (ESTs) T13618, and T45802. These primers were used to amplify a 2.24 kb fragment by PCR from genomic *Arabidopsis* DNA. The fragment was labeled with digoxigenin (DIG) using Boehringer Mannheims Dig-High Prime kit and protocol. This probe was used to screen a CD4-8 Landsberg *erecta* genomic library by plaque hybridization. About 2.7×10^5 recombinant phage were plated on the host *E. coli* LE392, grown overnight at 37 °C and screened. The hybridization temperature was 55 °C, everything else was done as described in the DIG Wash and Block Set protocol (Boehringer, Mannheim). Five positive clones were isolated of which four showed similar restriction patterns. One of them was subcloned into plasmid vector pBluescript SK +/− (Stratagene), transformed into DH5α competent cells (Gibco-BRL) and sequenced.

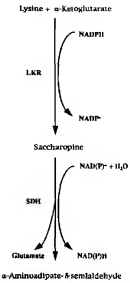


Figure 1. The first two steps of lysine catabolism in mammals and plants. LKR, lysine-ketoglutarate reductase; SDH, saccharopine dehydrogenase.

DNA sequencing and data analysis

DNA sequence analysis was carried out on both strands on an automatic sequencer (Model 377 and 373A) using the Ready Reaction FS Terminator sequencing kit and a 9600 Thermo-cycler (ABI, Applied Biosystems). Sequence data were analyzed using the Lasergene system (DNAstar, Wisconsin).

Isolation of total RNA

Whole *Arabidopsis* plants were frozen in liquid nitrogen and crushed in a mortar containing 4 ml of 1 M Tris-HCl pH 9.0 and 1% SDS. The extract was

transferred to a Sarstedt tube and 4 ml of a phenol/chloroform/isoamyl alcohol mixture (24:24:1 v/v/v) were added. The solution was vortexed and centrifuged at 12 000 × *g* for 10 min at room temperature. The supernatant was transferred to a new tube and 0.4 ml of a 2 M sodium acetate buffer, pH 5.2 and 8 ml of cold ethanol (70%) were added and the solution was kept on ice for 1 h. The nucleic acids were precipitated in a Sorvall centrifuge at 12 000 × *g* at 4 °C for 10 min and the supernatant discarded. The precipitate was dissolved in 2 ml of deionized sterile water and 2 ml of a 4 M solution of lithium acetate were added. After storage on ice overnight the RNA was precipitated at 12 000 × *g* at 4 °C for 10 min, washed with 2 ml of 70% ethanol, air-dried and dissolved in 0.4 ml of 10 mM Tris-HCl pH 7.5 in diethylpyrocarbonate (DEPC)-treated water. The RNA was stored at −70 °C until further analysis.

RT-PCR

RT-PCR was performed using a Perkin-Elmer kit. Total RNA (1 μg) from *Arabidopsis* was reverse-transcribed using oligo-dT as a primer. The LKR and SDH gene specific products were isolated using oligonucleotide primers, which were designed based on homologies of the genomic *LKR-SDH* DNA from *Arabidopsis* with the known coding sequences of the corresponding fungal proteins and ESTs T13618, T45802, and T04246. Overlapping clones were generated, subcloned into the pGEM-T (Promega), transformed into DH5α competent cells and sequenced.

Rapid amplification of cDNA ends (Race)

Isolation of the 5′ and 3′ cDNA ends was performed using the 5′ and 3′ Race systems for rapid amplification of cDNA ends (Gibco-BRL) according to the suppliers instructions. The reaction was started with 1 μg of total RNA. For the 5′ race the first and second gene specific primers were 5′-CAGCAGCCAATGAGGAAT-3′ and 5′-GCTGTCCAAGTCCGTGTAAGAAGTCAACA-3′, which are complementary to nucleotides 1262–1279 and 1093–1121 in Figure 3, respectively. Multiple bands were obtained after the first amplification. The largest band, which was 650 bp in length, was isolated and cloned into the pGEM-T vector (Promega). As the gene specific primer for the 3′ Race 5′-TCCTTGAAAGCAAACGTATAGAGAAGCACACT-3′ was used, which is identical to 5559–5590 in Figure 3. The 5′ and 3′ Race products were sequenced as described above.

Southern blotting

Total DNA was isolated from whole *Arabidopsis* plants and 10 μg were digested to completion with *Sst*I or *Nsi*I. The digests were loaded on 0.7% agarose gels, blotted onto Hybond N membrane (Amersham) and hybridized to a DIG-labeled probe corresponding to the 1.7 kb long *LKR* cDNA fragment. Hybridization, washing and detection procedures were as described in the DIG Wash and Block Set protocol from Boehringer Mannheim. The hybridization and washing temperatures chosen were 55 °C and 65 °C respectively.

Expression of Arabidopsis *LKR-SDH in* E. coli *cells*

The 3.2 kb long ORF coding for LKR-SDH was isolated by reverse transcription and subsequent PCR amplification using the oligonucleotides ATGAATTCAAATGGCCATGAGGAG and TCATTCTGCCTTCTCCATCAG, which are complementary to the 5′ and 3′ ends, respectively. The resulting PCR product was purified using the Promega PCR product purification kit and subjected to further amplifications using the oligonucleotides listed below:
1: TGAACCATGGCTTCAAATGGCCATGAGGAG
2: CATACCATGGCGAAAAAATCAGGTGTTT
3: TATGGTACCTCATTCAGGCTTCTCTTTTATCTC
4: TCTAGGTACCTCATTCTGCCTTCTCCATCAG

The complete LKR-SDH coding sequence was amplified using primers 1 and 4, which encompass the region between the 5′ and 3′ LKR-SDH coding sequence, respectively. The LKR and SDH coding sequencing were amplified separately using either primers 1 and 3 (LKR) or 2 and 4 (SDH), where primers 2 and 3 extend over nucleotides 3486–3503 and 3461–3481 (Figure 3), respectively. The primers added unique *Nco*I (primers 1 and 2) and *Kpn*I (primers 3 and 4) restriction sites (underlined) at the start codon and just past the stop codon of the gene, respectively. The generation of the *Nco*I sites resulted in the LKR region in a change of the second codon from asparagine to alanine and in the SDH cDNA in a change of the second codon from threonine to alanine. The PCR products were cloned into the *Nco*I and *Kpn*I restriction sites of the expression vector pBT430, a derivative of pET-3a [29] and transformed into BL21 (DE3)lysS cells (Novagene). Simultaneously cells were also transformed with the vector only.

The transformed cells were plated on LB medium containing ampicillin (100 μg/ml) and chloramphenicol (34 μg/ml) and grown overnight. Protein extracts from colonies resistant to antibiotics were subject to SDS-PAGE and analyzed for the relevant enzyme activity.

Preparation of extracts for enzyme activities and protein

Bacterial cultures (50 ml) were grown in LB media to an A_{600} of 0.6, IPTG was added to a final concentration of 1 mM and cells were grown for an additional 3 h. Then the cells were centrifuged (5000 × *g* for 10 min at 0–4 °C) and washed twice either with 100 mM phosphate buffer pH 7.0 (LKR) or with 100 mM Tris-HCl pH 8.5 (SDH) and resuspended in 2 ml of the relevant extraction buffer (see below). Extracts for the determination of LKR and SDH activities were prepared as described previously [15, 19] with some minor modifications. The LKR extraction buffer contained 100 mM phosphate buffer pH 7.0, 1 mM EDTA, 1 mM DTT and 15% glycerol. The SDH extraction buffer was composed of 100 mM Tris-HCl pH 8.5, 1 mM DTT, 1 mM EDTA, and 15% glycerol. The cell suspensions were frozen (−20 °C), thawed and sonicated at 0 °C for 1 min (30 s-1 min-30 s). The broken cells were centrifuged for 20 min at 10 000 × *g* at 0–4 °C and the resulting supernatant and pellet were subject to SDS-PAGE and analyzed for enzyme activities as described below. Protein concentration was determined according to Bradford [2], using BSA as a standard.

LKR specific activity

LKR specific activity was determined essentially as described [19], except for some minor modifications. The reaction mixture contained in 0.5 ml, 100 mM phosphate buffer pH 7.0, 20 mM lysine, 10 mM α-ketoglutarate, 0.1 mM NADPH, and 5–100 μg protein. Conversion of NADPH to NADP was followed at A_{340} at 30 °C. Aliquots were analyzed in the absence of lysine.

SDH specific activity

SDH specific activity was determined as described by Goncalves-Butrille *et al.* [15] with some minor modifications. The reaction buffer contained in 0.5 ml total volume, 20 mM saccharopine, 50 mM Tris-HCl pH 8.5, 20 mM NAD and 5 μg protein extract. Conversion of NAD to NADH was followed at A_{340} at 30 °C at the linear range. Control experiments were performed in the absence of saccharopine and activity was calculated by subtracting the values of the control assay from the values in the assay containing saccharopine.

Results

Gene isolation

The amino acid sequence for the fungal biosynthetic SDH proteins were used to search plant cDNA databases using the TBLASTN algorithm. We found two previously unidentified *Arabidopsis* ESTs (GenBank/EMBL accession numbers T13618 and T45802) that are homologous to the *Saccharomyces cerevisiae LYS9* gene. These ESTs were used to design primers and a 2.24 kb genomic fragment was amplified by PCR from genomic *Arabidopsis* DNA. The sequence similarity between the fungal glutamate-forming SDH and the isolated *Arabidopsis* fragment suggested that the latter contained coding sequences for SDH. Using the 2.24 kb fragment as a probe we screened a CD4-8 Landsberg *erecta* genomic library by plaque hybridization (see Materials and methods). One of the positive clones contained a nucleic acid fragment with regions that encoded a protein with domains homologous to fungal LKR (SDH-lysine-forming) and fungal SDH (SDH-glutamate-forming). During the sequencing of this DNA fragment another match with an *Arabidopsis* EST (GenBank/EMBL accession number T04246) was found in the 5′ LKR encoding region.

LKR-SDH *cDNA isolation*

Alignment between the fungal SDH proteins, the *Arabidopsis* ESTs and the genomic DNA fragment isolated from *Arabidopsis* allowed an approximate designation of the LKR and SDH coding sequences. Primers were designed and overlapping fragments of the corresponding cDNA were isolated from total *Arabidopsis* RNA by RT-PCR (Materials and methods). The sequences of the genomic DNA and cDNA fragments are shown in Figure 2.

Sequence analysis of the complete *LKR-SDH* cDNA revealed an ORF of 3.16 kb, which predicts a protein of 117 kDa. The deduced amino acid sequence from the cDNA indicates that LKR and SDH domains reside on one polypeptide in *Arabidopsis*. The observation that these two domains are linked, as has been reported for the purified corn LKR-SDH protein [15],

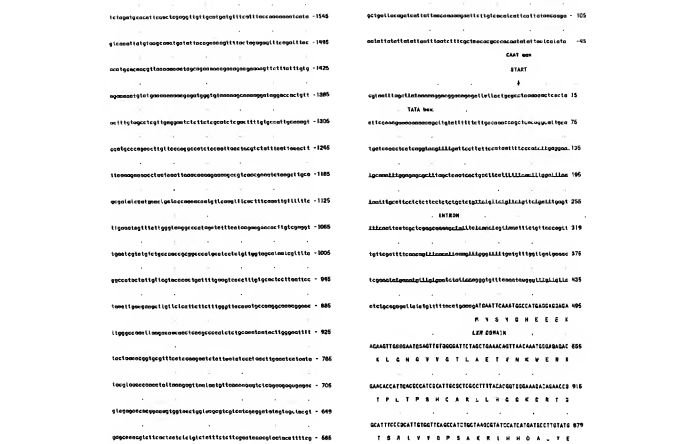


Figure 2. Nucleotide sequence and deduced amino acid sequence of the *A. thaliana LKR-SDH* gene Putative promoter elements and intron in 5′-flanking region are underlined. The cDNA ends are marked by arrows.

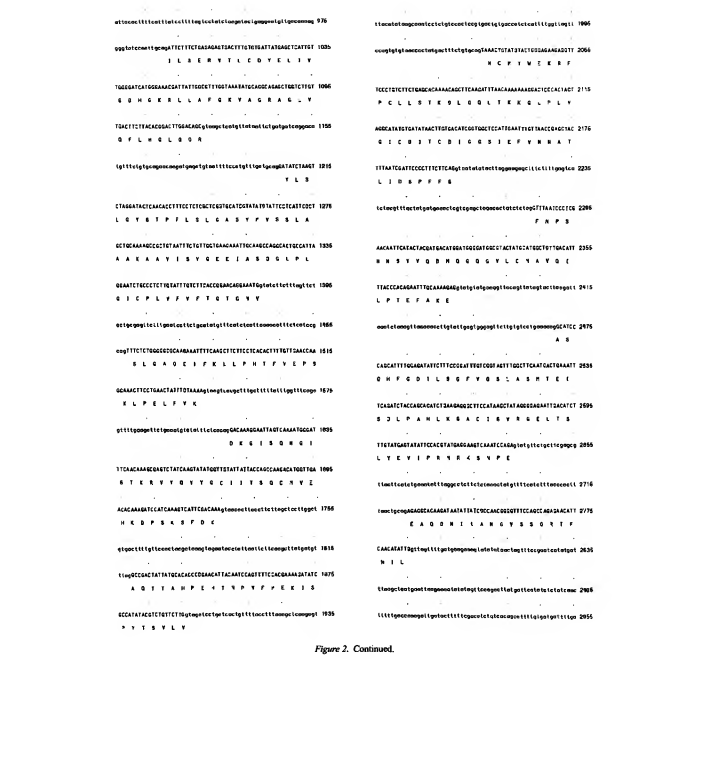


Figure 2. Continued.

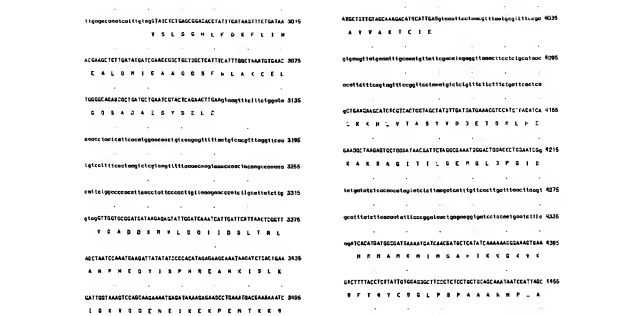


Figure 2. Continued.

provided additional confidence that this clone carried the *Arabidopsis LKR-SDH* gene.

Gene structure

A diagram of the *Arabidopsis LKR-SDH* gene structure is shown in Figure 3. The alignment of the genomic and

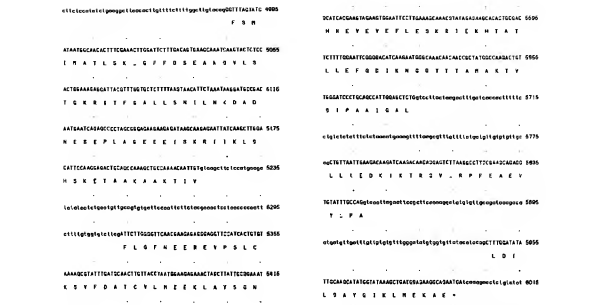


Figure 2. Continued.

cDNA sequences shows that the LKR-SDH apoprotein coding sequence is interrupted by 23 introns (Figures 2 and 3). The intron lengths range from 78 to 203 nucleotides with a majority lying in the range of 80–99 nucleotides; this agrees well with previous analyses of *Arabidopsis* introns. The intron sequences also fit very well with the plant splice junction consensus sequence (Table 1). Only one of the introns has a :GC at its 5′ end instead of a :GT; this divergence has been reported to occur at a low frequency in *Arabidopsis*. All introns have a > 58% AT content and all exceed the minimum required length of 66 nucleotides to ensure efficient splicing [4].

5′- and 3′-flanking regions

The 5′- and 3′-flanking regions of the *Arabidopsis LKR-SDH* gene were isolated using 5′ and 3′ Race systems (Materials and methods). The adapter primer and the gene specific primer chosen for amplification

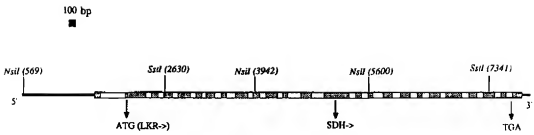


Figure 3. Restriction map and exon-intron pattern of the *A. thaliana LKR-SDH* gene. Introns (□), exons (■), 5′- and 3′-flanking regions (□), and chromosome (▬).

Table 1. Percentage occurrence of the consensus nucleotides of *Arabidopsis* introns at the 5′ and 3′ splice sites.

	5′ splice site							3′ splice site					
position	−2	−1	+1	+2	+3	+4	+5	−5	−4	−3	−2	−1	+1
consensus	A	G	G	T	A	A	G	T	G	C	A	G	G
% occurrence	62	71	100	96	79	33	50	24	54	75	100	100	50

always bridged a region containing an intron in the genomic sequence to ensure that no contaminating DNA was amplified.

Amplification of the 5′ Race product led to the production of several bands, probably due to incomplete reverse transcriptase reactions. The largest band was gel isolated, analyzed with appropriate restriction enzymes, cloned and sequenced. Transcription most likely starts at a CTA sequence (Figure 2); upstream from the putative transcription start point are TATAAA and CAAT sequences. The TATAAA sequence begins at −33 and the CAAT sequence begins at −63. The position of translation initiation, the ATG codon (position 1 of the deduced amino acid sequence), indicates that the mRNA contains a 108 base long 5′ leader sequence, which is interrupted by an 352 base long intron (Figure 2).

Multiple ATGs that are out of frame with the *LKR-SDH* coding sequence, were identified in the 5′-untranslated region (positions 74, 451 and 461) (Figure 2). The ability of eukaryotic ribosomes to initiate translation requires the AUG to reside in the consensus sequence ANNAUGN or NNNA UGG [22, 23]. The AUG's in the 5′ untranslated leader of the *Arabidopsis* mRNA are not flanked by these consensus initiator sequences. The functions, if any, of these upstream AUGs is unknown.

The 5′-upstream region was analyzed for other consensus sequences. The *Opaque-2* gene product transactivates expression of the 22 kDa α-zein genes in maize endosperm and evidence exists which suggests that LKR could also be under *Opaque-2* control [3]. Therefore, we analyzed the 5′ leader for the consensus sequence of the Opaque 2-binding site GATGAPyPuTGPu [24]. No match between this consensus sequence and the 5′-flanking region was found. Apparently the lysine degradative pathway operates in the seeds of various higher plants and might be confined to them, hence we looked for homology with the Sph box CATGCATG, a *cis*-regulatory element conferring seed-specific expression [11, 30]. No match was found with this sequence either. Sequence analysis for other binding sites of known plant transcription factors [20] did not show any perfectly conserved binding sites.

The 3′ terminus of the cDNA sequence was amplified using the 3′ Race system, resulting in the formation of only one product, which subsequently was cloned and sequenced. The 3′-untranslated sequence extends 90 bases past the stop codon. A poly(A)$^+$ addition signal resembling the animal consensus sequence AATAAA is not seen at the 3′ terminus of the cDNA sequence (Figure 2). In the case of the *LKR-SDH* gene it might be that the poly(A) addition signal is different from the consensus. Although some plant genes do have the unaltered AATAAA motif, plants seem to be more divergent in this motif and other sequences up and downstream of the poly(A) cleavage site might compensate for lack of an AATAAA sequence [16]. Another possibility would be that reverse transcription during the 3′ Race started from an internal run of A

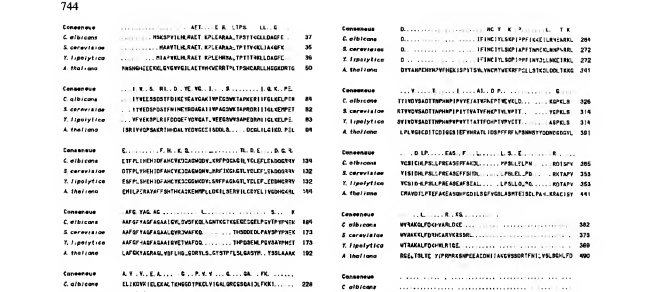


Figure 4. Comparison of the deduced amino acid sequences of three fungal genes encoding SDH (lysine-forming) with the *A. thaliana* LKR. The alignment was created by the program PileUp (GCG Package). Amino acid residues that are identical between at least one of the fungal proteins and the plant protein appear in the consensus.

residues in the mRNA and not from the true poly(A) tail. Although a consensus poly(A) signal is present 261 bases past the stop codon, it does not seem likely to be the true polyadenylation signal, since during the 3′ Race only one band was generated and subsequent amplifications using nested gene specific primers did not lead to the appearance of additional bands (not shown). The same was true when the amount of the oligonucleotide primer used for reverse transcription was decreased (not shown).

Southern blot analysis

Southern blot analysis was used to determine the number of genes encoding LKR-SDH in *A. thaliana* (not shown). Total DNA was digested with *Sst*I and *Nsi*I and hybridized with a digitonin-labeled cDNA probe corresponding to the first 1.7 kb of the LKR-coding sequence (see Materials and methods). The expected band lengths can be deduced from the diagram in Figure 3. Digestion with *Sst*I yielded a 4.7 kb and a 4.2 kb band. The latter band originates from digestion of a third *Sst*I site (not shown), which is present 4.2 kb upstream to the first *Sst*I site in the *LKR* gene. Digestion with *Nsi*I yielded two bands of 3.3 and 1.6 kb as expected. Under the conditions applied it thus appears that the *LKR-SDH* gene is present as a single copy in *A. thaliana*.

Deduced amino acid sequence and sequence comparisons

Pairwise comparisons between the deduced amino acid sequence of the *Arabidopsis* LKR-SDH and the four fungal SDH protein sequences were made with the computer program Bestfit (GCG package, Wisconsin). The deduced LKR protein from *Arabidopsis* shows an identity of about 25% with the fungal proteins. The homology increases to about 50% with the inclusion of conservative substitutions (data not shown). Alignment

of the deduced *Arabidopsis* SDH protein sequence shows a sequence identity of 37% and similarity of 57% to the glutamate-forming SDH from *S. cerevisiae*. Optimal alignments between the *Arabidopsis* LKR and the three fungal lysine-forming SDHs were made with the program PileUp (GCG package), (Figure 4) and between the *Arabidopsis* SDH and the yeast SDH with Bestfit (Figure 5). The alignments of the LKR and SDH homologues reveal several stretches of conserved residues that may be important for the function of this enzyme (Figures 4 and 5).

Lack of a chloroplast transit peptide sequence

Enzymes involved in lysine biosynthesis have been located in the chloroplasts of plants [5] and many of the enzymes have been shown to be synthesized in the form of preproteins [13, 14, 17, 32]. The preproteins have amino-terminal extensions, chloroplast transit peptides (CTPs), which direct them from the cytoplasm into the chloroplast and which are subsequently removed from the protein upon entering the latter [21]. The *Arabidopsis LKR-SDH* gene studied in this work encodes a protein that appears to lack an N-terminal chloroplast targeting sequence, since it disagrees with at least three observations made by von Heijne *et al.* in a comparison of 26 CTPs [31]. The second amino acid is not an alanine, there are 4 charged groups in the first 10 residues (Glu-7, 8, 9 and Lys-10). Serine, which is present at a level of 20% in CTPs, is not enriched in the first 100 residues of the *LKR-SDH* gene (Figure 2). Furthermore, homology to the fungal SDH (lysine-forming) protein begins at amino acid 21 of *Arabidopsis* LKR-SDH. Thus it appears unlikely that the *Arabidopsis* protein is targeted, implying that at least the first two steps of this lysine degradative pathway occur in the plant cell cytosol.

Expression of LKR-SDH in E. coli

Fragments encoding either the LKR domain or the SDH domain or the complete bifunctional protein were generated using PCR primers with appropriate restriction sites. The amplified fragments were digested, ligated to a prokaryotic expression vector and transformed into *E. coli* (see Materials and methods). The LKR domain alone or the complete protein coding sequence did not lead to the synthesis of detectable protein or enzyme activity (not shown). The failure to express these proteins was not due to mutations introduced by the amplification process. Efforts are now underway to express the bifunctional protein and the LKR domain in an eukaryotic expression system. In contrast, we were able to successfully express the SDH domain in *E. coli* leading to high protein levels and a high specific activity (Figures 6 and 7). The SDH coding region encompasses 1.4 kb on the cDNA clone, which predicts a protein of 52 676 Da. Extracts from IPTG induced cells that were transformed with the vector carrying the 1.4 kb insert were analyzed by SDS-PAGE and a protein at the expected size was overproduced in these cells (Figure 6). Separation of the cell extracts into its supernatant (lane C) and pellet (lane D) fraction shows that substantial amounts of protein are present in both of them. No band of similar intensity was present in uninduced cells that carry the vector + insert, or

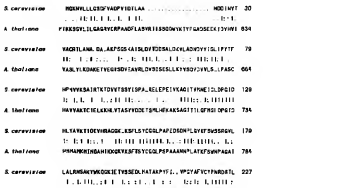


Figure 5. Comparison of the deduced amino acid sequence of the *S. cerevisiae* SDH (glutamateforming) and the *A. thaliana* SDH. The alignment was created by the program BestFit (GCG Package). Identical residues are designated by bars, conserved substitutions by dots.

Figure 6. SDS-PAGE of protein extracts. Cells were grown and protein extracts prepared as described in Materials and methods and subjected to SDS-PAGE. Size markers (lane A). Extracts from cells carrying vector + SDH insert (lanes B, C, D). Uninduced cells (lane B). IPTG-induced extract supernatant (lane C). IPTG-induced extract pellet (lane D). IPTG-induced empty vector (lane E).

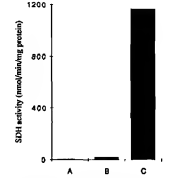


Figure 7. SDH activity in bacterial extracts. SDH activity was assayed as described in Materials and methods. The reaction was started by the addition of 5 μg of protein of the bacterial extract. Assay of extracts from cells without (A) or with SDH cDNA insert (B+C), assayed in the presence (A+C) or absence (B) of saccharopine.

induced cells that carry an empty vector (lanes B and E, respectively).

SDH activity was measured in the soluble fraction of the bacterial extracts (Figure 7). As expected no SDH activity was observed in extracts from cells transformed with an empty vector (column A). Extracts from cells containing the SDH cDNA insert converted substantial amounts of NAD+ to NADH (column C). The reaction was specific for SDH in that no significant activity was observed in the absence of the SDH substrate saccharopine (column B).

Similar to the maize and mammalian enzyme, activity of the *Arabidopsis* SDH increases from pH 6.0 to 9.0 and retains at these pH values 10% activity when NAD+ is replaced by NADP+ (data not shown).

Discussion

As a first approach towards understanding the physiological role of the lysine breakdown pathway in plants, we have isolated and characterized a gene encoding LKR-SDH from *A. thaliana*. The gene encoding the LKR-SDH protein covers about 6.2 kb of the *Arabidopsis* genome. Sequence analysis of the cDNA revealed an ORF of 3.16 kb, which predicts a protein of 117 kDa. The alignment of the genomic and cDNA sequences shows that the LKR-SDH gene is interrupted by 24 introns. They are of small size, as expected for *Arabidopsis* introns, and are predominantly of the pyrimidine-rich class. Upstream from the putative transcription start point, TATAAA and CAAT sequences were found at positions consistent with those of functional TATA and CAAT boxes reported previously for other eukaryotic genes [7]. Although existing data implicates the Opaque 2 transactivator being involved in the regulation of expression of LKR in maize endosperm [3], a search in the 5′-flanking region of the *Arabidopsis* gene did not reveal an *opaque 2* regulatory element. It is possible that the sequence of the Opaque 2 binding site diverges in the present case from the known consensus. Alternatively, regulation of LKR-SDH in different plants may vary or Opaque 2 may affect LKR-SDH indirectly, for example by the induction of the synthesis of an intermediary regulatory molecule. LKR activity appears to be restricted to the seeds of plants [18, Falco *et al.*, unpublished results]. Hence, we analyzed the 5′ flanking region for a Sph box, a sequence element, which has been shown to be involved in the seed specific expression of several plant genes. No sequence resembling the Sph box was detected. A functional analysis of the 5′-transcribed region will be needed to further elucidate the regulation of expression of LKR-SDH in *Arabidopsis thaliana*.

The deduced LKR and SDH amino acid sequences from *Arabidopsis* show an identity of about 25% and 37% and a similarity of 50% and 57%, respectively, to the corresponding fungal proteins. Although LKR and SDH reside on one polypeptide in *Arabidopsis*, we were able to functionally express SDH separately from the LKR domain in bacteria. This activity was similar in its biochemical characteristics to those of the corresponding enzyme purified from maize [15]. We have so far been unable to express either the LKR

domain or the entire LKR-SDH protein in *E. coli*. In the amino acid sequence of the *Candida albicans* SDH (lysine-forming) protein, residues 194–224 have been indicated as being important in NADH binding [12]. Nine residues in this stretch of amino acids match a fingerprint determined by Wierega *et al.* [33], the most important being three glycines at positions 6, 8, and 11 and an acidic amino acid at the last position of the peptide. In the case of a NADPH-binding site the latter would be expected to be exchanged for a hydrophobic residue. An 'ADP-binding fold' or 'fingerprint' was not found in either the LKR or in the SDH domain of *Arabidopsis* as such. In some cases, however, variations in this fingerprint have been reported [12, 33].

There are about 200 amino acid residues in the *Arabidopsis* LKR-SDH protein between the regions homologous to fungal SDH (lysine-forming) and fungal SDH (glutamate-forming), which is suggestive of an intermediary or 'spacer' region. However, to define the role of this region, a functional analysis of the LKR domain and isolation of other bifunctional *LKR-SDH* genes are necessary.

In contrast to the lysine biosynthetic pathway, which appears to operate in the plastids of plant cells, our results suggest that the *Arabidopsis* protein is not targeted to the chloroplast, implying that at least the first two steps of this lysine degradative pathway occur in the cytosol of plants. A gene encoding a chloroplast-targeted isoform of the protein does not seem to exist, since standard Southern blot analysis using *Arabidopsis LKR* cDNA as a hybridization probe, suggested that there is a single copy of the bifunctional LKR-SDH protein.

The results presented here have practical implications. It has been shown that LKR-SDH participates in one of the major lysine breakdown pathways. Function of this pathway interferes with the efficiency of lysine accumulation in seeds of transgenic crop plants, which were engineered to synthesize high levels of lysine ([9], Falco *et al.* unpublished results). Inactivation of LKR-SDH through genetic engineering therefore might be a feasible way to increase lysine accumulation on the one hand and avoid formation of undesired lysine breakdown products on the other hand. As a first step to accomplish this, we have used the *Arabidopsis LKR-SDH* gene to obtain the corresponding genes from soybean and corn. Furthermore, LKR has been shown to be also an important enzyme in mammalian cells. The human genetic disease familial hyperlysinemia is caused by the accumulation of lysine in mitochondria, which is caused by a defect in production of the LKR-SDH enzyme and hence results in a decrease or absence of lysine catabolism [25]. This study should simplify the isolation of the genes for these catabolic enzymes from animal sources, as it has for other plants.

Acknowledgement

We thank S. Stack for excellent technical assistance.

References

1. Arruda P, Da Silva WJ: Lysine-ketoglutarate reductase activity in developing maize endosperm. Phytochemistry 22: 2687–2689 (1983).
2. Bradford M: A rapid and sensitive method for quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72: 248–254 (1976).
3. Brochetto-Braga M, Leite A, Arruda P: Partial purification and characterization of lysine-ketoglutarate reductase in normal and *opaque-2* maize endosperms. Plant Physiol 98: 1139–1147 (1992).
4. Brown JWS, Smith P, Simpson CG: *Arabidopsis* consensus intron sequences. Plant Mol Biol 32: 531–535 (1996).
5. Bryan JK: Advances in the biochemistry of amino acid biosynthesis. In: Miflin BJ, Lea PJ (eds) Intermediary Nitrogen Metabolism, pp. 161–196. The Biochemistry of Plants, vol. 16. Academic Press, New York (1990).
6. Bryan JK: Synthesis of the aspartate family and branched-chain amino acids. In: Miflin BJ (ed) Amino Acids and Derivatives, pp. 403–452. The Biochemistry of Plants, vol. 5. Academic Press, New York (1980).
7. Buchner P: Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212: 563–578 (1990).
8. Cohen NG, Saint-Girons I: Biosynthesis of threonine, lysine and methionine. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology, vol I, pp. 429–444. American Society for Microbiology, Washington, DC (1987).
9. Falco SC, Guida T, Locke M, Mauvis J, Sanders C, Ward RT, Webber P: Transgenic canola and soybean seeds with increased lysine. Bio/technology 13: 577–581.
10. Feller A, Dubois E, Ramos F, Pierard A: Repression of the genes for lysine biosynthesis in Saccharomyces cerevisiae is caused by limitation of Lys14-dependent transcriptional activation. Mol Cell Biol 14: 6411–6418 (1994).
11. Fujiwara T, Beachy RN: Tissue-specific and temporal regulation of a β-conglycinin gene: roles of the RY repeat and other *cis*-acting elements. Plant Mol Biol 24: 261–272 (1994).
12. Garrad R, Schmidt TM, Bhattacharjee JK: Molecular and functional analysis of the *LYS1* gene of *Candida albicans*. Infect Immun 62: 5027–5031 (1994).
13. Ghislain MV, Frankard V, Jacobs M: Dihydropicolinate synthase of *Nicotiana sylvestris*, a chloroplast-localized enzyme of the lysine pathway. Planta 180: 480–486 (1990).
14. Ghislain M, Frankard V, Vandenbossche D, Matthews F, Jacobs M: Molecular analysis of the aspartate kinase-homoserine

dehydrogenase gene from *Arabidopsis thaliana*. Plant Mol Biol 24: 835–851 (1994).
15. Goncalves-butruille M, Szajner P, Torigoi E, Leite A, Arruda P: Purification and characterization of the bifunctional enzyme lysine-ketoglutarate reductase-saccharopine dehydrogenase from maize. Plant Physiol 110: 765–771 (1996).
16. Joshi CP: Putative polyadenylation signals in nuclear genes of higher plants: a compilation and analysis. Nucl Acids Res 15: 9627–9640 (1987).
17. Kaneko T, Hashimoto T, Kumpaisal R, Yamada Y: Molecular cloning of wheat dihydropicolinate synthase. J Biol Chem 265: 17451–17455 (1990).
18. Karchi H, Miron D, Ben-Yaacov S, Galili G: The lysine-dependent stimulation of lysine catabolism in tobacco seed requires calcium and protein phosphorylation. Plant Cell 7: 1963–1970 (1995).
19. Karchi H, Shaul O, Galili G: Lysine synthesis and catabolism are coordinately regulated during tobacco seed development. Proc Natl Acad Sci USA 91: 2577–2581 (1994).
20. Katagiri F, Chua NH: Plant transcription factors: present knowledge and future challenges. Trends Genet 8: 22–27 (1992).
21. Koegstra K, Olsen LJ, Theg SM: Chloroplastic precursors and their transport across the envelope membranes. Plant Mol Biol 40: 471–501 (1989).
22. Kozak M: Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. Proc Natl Acad Sci USA 92: 2662–2666 (1995).
23. Kozak M: The scanning model for translation: an update. J Cell Biol 108: 229–235 (1989).
24. Lohmer S, Maddaloni M, Motto M, Di Fonzo N, Hartings H, Salamini F, Thompson RD: The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. EMBO J 10: 617–624 (1991).
25. Markovitz PJ, Chuang DT: Familial hyperlysinemias: purification and characterization of the bifunctional aminoadipic semialdehyde synthase with LKR and SDH activities. J Biol Chem 259: 11643–11646 (1984).
26. Mazelis M. Amino acid metabolism. In: Miflin BJ (ed) Amino Acids and Derivatives, pp. 541–567. The Biochemistry of Plants, vol. 5. Academic Press, New York (1980).
27. Mills WR, Lea PJ, Miflin BJ: Photosynthetic formation of the aspartate family of amino acids in isolated chloroplasts. Plant Physiol 65: 1166–1172 (1980).
28. Ogawa H, Fujioka M: Purification and characterization of saccharopine dehydrogenase from baker's yeast. Biol Chem 253: 3666–3670 (1978).
29. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW: Use of T7 polymerase to direct expression of cloned genes. Meth Enzymol 185: 60–89 (1990).
30. Uno T, Yamaguchi-Shinozaki K, Mitsukawa N, Shibata D, Shinozaki K: Molecular cloning and characterization of a gene that encodes a MYC-related protein in *Arabidopsis*. Plant Mol Biol 32: 571–576 (1996).
31. Von Heijne G, Steppuhn J, Hermann RG: Domain structure of mitochondrial and chloroplast targeting peptides. Eur J Biochem 180: 535–545 (1989).
32. Weisemann JM, Matthews BF: Identification and expression of a cDNA from *Daucus carota* encoding a bifunctional aspartokinase-homoserine dehydrogenase. Plant Mol Biol 22: 301–312 (1993).
33. Wierenga RK, Terpstra P, Hol WMG: Prediction of the occurrence of the ADP-binding $\beta\alpha\beta$-fold in proteins, using an amino acid sequence fingerprint. J Mol Biol 187: 101–107 (1986).
34. Yuan J-W, Fournier P, Declerck N, Chasles M, Gaillardin C: Overlapping reading frames in the *LYS5* locus in the yeast *Yarrowia lipolytica*. Mol Cell Biol 10: 4795–4806 (1990).

Evidence Appendix B

Rule 132 Declaration of Dr. Carl Falco dated August 24, 2000. (Note: The original declaration can be found in the file of Application No. 08/823,771.)

A copy of this declaration accompanies the Response After Final submitted on February 4, 2008 and was entered by the Examiner, Office Communication dated March 5, 2008.

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

IN THE APPLICATION OF:

SAVERIO C. FALCO
SHARON J. KEELER
JANET A. RICE

CASE NO.: BB-1037-D

APPLN. NO.: 08/823,771

GROUP ART UNIT: 1638

FILED: MARCH 24, 1997

EXAMINER: E. MCELWAIN

FOR: CHIMERIC GENES AND METHODS FOR INCREASING THE LYSINE AND THREONINE CONTENT OF THE SEEDS OF PLANTS

Date: AUGUST 24, 2000

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

Declaration of Dr. Carl Falco Pursuant to 37 CFR §1.132

I, Saverio Carl Falco, am a citizen of the United States of America, residing at 1902 Miller Road, Arden, Delaware 19810, United States of America, and I declare as follows:

1. I am one of the above-identified inventors named in this application. I am a graduate of Rutgers University of New Brunswick, New Jersey with a B.A. degree granted in 1971 with high honors and distinction in physics. I received a Ph.D. in 1977 from the University of Chicago in biochemistry and molecular biology. From 1977 to 1981 I was a National Institutes of Health postdoctoral fellow at the Massachusetts Institute of Technology. I have been employed by E. I. du Pont de Nemours and Company since 1981 directing and conducting research in plant genetic engineering.

2. I have reviewed the Office Action dated April 25, 2000. I am aware that this declaration is being submitted to address the concerns set forth on page 4 and 5 of the Office Action that "the specification does not disclose any plants that comprise the claimed two gene fragments that result in the claimed increase in lysine relative to a

I HEREBY CERTIFY THAT THIS PAPER IS BEING DEPOSITED WITH THE UNITED STATES POSTAL SERVICE WITH SUFFICIENT POSTAGE AS FIRST CLASS MAIL IN AN ENVELOPE ADDRESSED TO: ASST. COMMISSIONER FOR PATENTS, WASHINGTON, D.C. 20231, ON THIS DATE.

8/24/00
Date

R. Nanette Marcus
R. Nanette Marcus

plant that does not comprise said two gene fragments. In addition, the specification fails to provide guidance with regard to the choice of subfragments that will result in the antisense inhibition or cosuppression of LKR."

3. At the outset, it is noted that many components of the process of plant genetic engineering, e.g. construction of chimeric genes for expression in plant cells, or for blocking expression of endogenous genes, transformation of plants, have become routine for those skilled in the art. Notwithstanding this, what follows is intended to show that one of ordinary skill in the art could follow the teachings of the instant application to practice the claimed invention without engaging in undue experimentation.

4. First, the rationale for combining the nucleic acid fragments of the invention are clearly disclosed in the specification. It was shown, for the first time, that accumulation of excess free lysine in plant seeds, accomplished via expression of lysine insensitive DHDPS, is accompanied by breakdown of free lysine and accumulation of intermediates in the breakdown pathway such as saccharopine. Thus, there was a clear incentive to reduce the loss of excess lysine due to catabolism.

5. Second, methods were provided to prevent lysine catabolism through reduction in the activity of the enzyme lysine ketoglutarate reductase (LKR), which catalyzes the first step in lysine breakdown. This can be accomplished by introducing a mutation in the plant gene that encodes LKR that reduces or eliminates enzyme function. Such mutations can be identified by screening mutants for lysine over-producer lines that do not accumulate the lysine breakdown products, saccharopine and α-amino adipic acid. Alternatively, the first nucleic acid fragments containing plant LKR cDNAs were disclosed. The nucleotide sequences of these fragments make it straightforward to isolate LKR nucleic acid fragments from any plant desired (see point 6 below). Chimeric genes for expression of antisense LKR RNA or for cosuppression of LKR in the seeds of plants can then be created. The chimeric LKR gene can be linked to chimeric genes encoding lysine insensitive AK and DHDPS and all introduced into plants via transformation simultaneously, or the chimeric LKR gene or mutant LKR gene can be brought together with chimeric genes encoding lysine insensitive AK and DHDPS by crossing plants to create hybrids carrying two or more of the genes (see below).

6. Third, examples of all of the nucleic acid fragments of the invention were provided in the specification of the subject case. In the case of the bifunctional protein lysine ketoglutarate reductase (LKR)/saccharopine dehydrogenase (SDH), two plant nucleic acid fragments (SEQ ID NOS:102 and 103) containing cDNA derived

from the plant *Arabidopsis thaliana* were provided in the present patent application. In the application it was stated that full length cDNAs encoding plant LKR plus saccharopine dehydrogenase (SDH) or genomic DNAs containing the entire LKR/SDH gene can be readily identified by hybridization to labelled cDNA fragments of SEQ ID NO:102: or SEQ ID NO:103: and thus isolated. This was, in fact, accomplished and is described in Epelbaum, S., McDevitt, R. and Falco, S. C., (1997) "Lysine-ketoglutarate reductase and saccharopine dehydrogenase from *Arabidopsis thaliana*: nucleotide sequence and characterization", Plant Mol. Biol. 35, 735.

The availability of the Arabidopsis LKR/SDH gene made it straightforward for us, as it would be for anyone skilled in the art, to isolate other plant LKR/SDH genes. Degenerate oligonucleotides were designed based upon highly conserved regions of the deduced amino acid sequence of plant and fungal proteins and used to amplify soybean and corn LKR/SDH cDNA fragments. Near full-length cDNAs for soybean and corn LKR/SDH were then isolated using 5' RACE and hybridization to cDNA libraries. LKR/SDH nucleic acid fragments were isolated from several other plant species including wheat and rice by identifying EST sequences homologous to the already known plant LKR/SDH sequences.

7. Fourth, there is a description of how to use these nucleic acid fragments to practice the invention. In the case of LKR/SDH, the availability of plant LKR/SDH genes made it possible to block expression of the LKR/SDH gene in transformed plants via antisense inhibition or cosuppression. It was stated in the Office Action on page 4 that antisense inhibition and cosuppression of a gene in a plant is unpredictable. This is true only in the sense that every transformant does not produce the desired phenotype. But one skilled in the art is well aware of this and designs the experiment in a way that many transformants are obtained and screened for the desired phenotype.

My own experience with cosuppression methodology in plants, as well as my knowledge of the work of my colleagues, and research work in the broader scientific community, indicates that this method is reliable and predictable. The use of cosuppression to block expression of several different genes in several different plants has been achieved [illegible] successfully at DuPont.

Specifically in the case of LKR/SDH, cosuppression has been used to block expression with the first gene fragment and promoter combination tested, which hardly represents undue experimentation (see point 10 below).

8. It is stated on page 5 of the Office Action that "De Luca teaches that modifying metabolic pathways by transforming plants with genes that control steps of the pathway is highly unpredictable and often the desirable results are impossible to achieve." This may be true in cases where not enough is known about the metabolic pathway, but in the case of the lysine biosynthetic and catabolic pathways, it has been demonstrated how to increase production of lysine via modification of the biosynthetic pathway using lysine insensitive DHDPS and AK, and shown that accumulation of free lysine in seeds is also controlled by catabolism of lysine. We teach that blocking the first step in lysine catabolism will lead to increased accumulation of lysine and this is, in fact, what we have observed as described below.

9. The corn LKR/SDH cDNA sequence was used to identify transposon mutations in the endogenous corn LKR/SDH gene via PCR screening of a library of corn lines containing Robertson's Mutator transposon insertions. The precise location of Mutator insertions into the LKR/SDH gene was determined by sequencing of genomic DNA from individual mutants. An insertion mutation located in an exon in the LKR domain of the gene was chosen for further study. Southern blot analysis of corn genomic DNA indicated that corn contains only one LKR/SDH gene. Since an insertion mutation is expected to block function of the gene, it was anticipated that such a mutation would be recessive. One fourth of the progeny seed from a selfed corn ear with such a mutation segregating would be expected to be homozygous for the mutation. It was observed that approximately one fourth of such seed exhibited a higher level of free lysine than normal (5 to 15 fold higher) without the increase in the lysine catabolite saccharopine that is seen when free lysine is increased via expression of lysine insensitive DHDPS. It was concluded that knocking out LKR/SDH, by itself, was able to increase seed lysine content in corn seeds.

The LKR/SDH Mutator insertion line was crossed by a transgenic line that accumulates excess free lysine due to expression of lysine insensitive DHDPS and AK. In this cross two genetic loci that affect lysine accumulation, one of which is recessive (the LKR/SDH Mutator insertion) and one of which is semi-dominant (the lysine insensitive DHDPS and AK trangene locus), are segregating. Single seeds were analyzed for lysine and saccharopine content. The most striking observation from this experiment is that the highest lysine containing seeds have low levels of saccharopine (see figure). The low saccharopine level indicates that these seeds are homozygous for the LKR/SDH Mutator insertion, while the high lysine level indicates that they carry the lysine insensitive DHDPS and AK trangene locus. The level of lysine accumulation is considerably higher (2-3 fold) than the level provided by the

DHDPS and AK trangene locus alone. Thus, this experiment demonstrates that an increase in the accumulation of lysine, accompanied by a reduction in accumulation of lysine catabolites can be accomplished by combination of lysine overproduction brought about by expression of lysine insensitive DHDPS + AK and reduction of lysine catabolism by blocking expression of LKR/SDH, as we taught in the patent application. These results show that the concern stated in the Office Action on page 5 that "modifying metabolic pathways ... is highly unpredictable and often the desirable results are impossible to achieve" is unfounded in this particular case.

10. As indicated above, LKR/SDH expression has been blocked in corn via cosuppression. To accomplish this a chimeric gene designed for cosuppression of LKR was constructed by linking a 1268 bp LKR/SDH gene fragment, which included the LKR coding domain, to the corn endosperm 27 kD zein promoter and 10 kD zein 3' untranslated region. This chimeric gene was introduced into corn by particle-gun mediated transformation. Of 72 transformation events that were regenerated into plants and produced seed, 13 had seeds with a greater than four fold increase in free lysine. This is a typical frequency for cosuppression events. Since the transformed plants were out-crossed, the transgenic locus must be dominant or there would not have been any observable phenotype. This is expected from a cosuppression transgene, and is an advantage over knock-out mutations like the LKR/SDH Mutator insertion described above.

Some of the LKR cosuppression transformants have been carried forward for further testing. An event that has continued to show the increased free lysine phenotype for several generations and behaves genetically as a single locus transgene insertion has been selected for crossing to the transgenic line that accumulates excess free lysine due to expression of lysine insensitive DHDPS and AK. Results from that experiment are not yet available, but the expectation is that seeds carrying both transgene loci will have higher lysine levels than either parent, as was observed in the LKR Mutator insertion cross described above. In addition, co-transformation experiments in which the chimeric gene designed for cosuppression of LKR described above has been combined with a chimeric gene for expression of lysine insensitive DHDPS and introduced into corn by particle-gun mediated transformation are in progress. This is expected to yield transformants that produce seeds with the high lysine level observed in the LKR Mutator insertion cross by lysine insensitive DHDPS and AK, but with both chimeric genes at a single genetic locus, which is highly desirable for corn breeding.

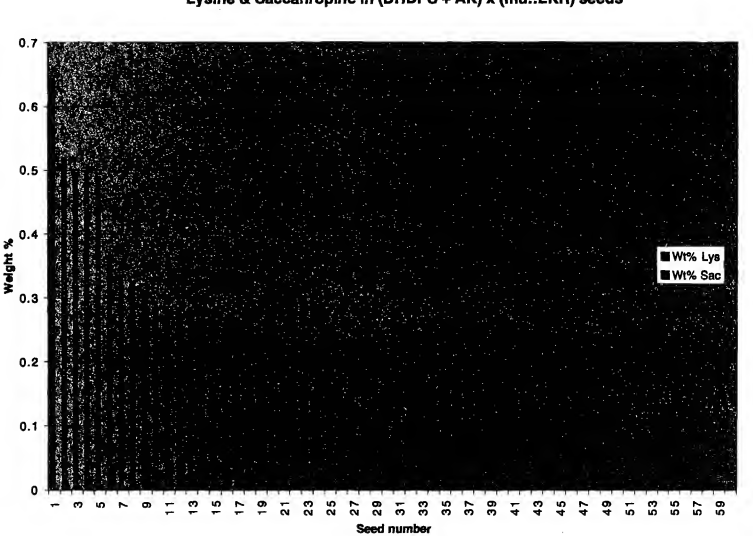
In summary, all of the elements of the claimed invention were provided in the patent application. The teachings in this case are in the public domain, due to the issuance of U. S. Patent 5,773,691 of which the instant application claims priority as a divisional application.. One skilled in the art can take these elements, as discussed above, and practice the invention without undue experimentation.

I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements are made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Saverio Carl Falco

8/24/00

Date


Lysine & Saccahropine in (DHDPS + AK) x (mu::LKR) seeds
Weight %
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Wt% Lys
Wt% Sac
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Seed number

Evidence Appendix C

Rule 132 Declaration Declaration of Dr. Carl Falco dated February 16, 2001 (Note: the original declaration can be found in the filed of Application No. 08/823,771.)

A copy of this declaration accompanies the Response After Final submitted on February 4, 2008 and was entered by the Examiner, Office Communication dated March 5, 2008.

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

IN THE APPLICATION OF:

SAVERIO C. FALCO
SHARON J. KEELER
JANET A. RICE

CASE NO.: BB-1037-D

APPLN. NO.: 08/823,771

GROUP ART UNIT: 1638

FILED: MARCH 24, 1997

EXAMINER: E. MCELWAIN

FOR: CHIMERIC GENES AND METHODS FOR INCREASING THE LYSINE AND THREONINE CONTENT OF THE SEEDS OF PLANTS

Date: FEBRUARY 16, 2001

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

Declaration of Dr. Carl Falco Pursuant to 37 CFR §1.132

I, Saverio Carl Falco, am a citizen of the United States of America, residing at 1902 Millers Road, Arden, Delaware 19810, United States of America, and I declare as follows:

1. I am one of the above-identified inventors named in this application. I am a graduate of Rutgers University of New Brunswick, New Jersey with a B.A. degree granted in 1971 with high honors and distinction in physics. I received a Ph.D. in 1977 from the University of Chicago in biochemistry and molecular biology. From 1977 to 1981 I was a National Institutes of Health postdoctoral fellow at the Massachusetts Institute of Technology. I have been employed by E. I. du Pont de Nemours and Company since 1981 directing and conducting research in plant genetic engineering.

2. I have reviewed the Office Action dated November 22, 2000. I am aware that this declaration is being submitted to address the concerns set forth on page 3 of the Office Action that the "Declaration of Falco teaches use of a bifunctional LKR/SDH gene to identify mutants produced by transposon mutagenesis. This plant does not contain a foreign LKR gene. In addition, the Declaration of Falco teaches of a combination DHDPS gene without an AK gene. Thus, the Declaration of Falco

does not teach a plant with a foreign LKR gene and a foreign DHDPS gene . . . it remains unpredictable what the results would be of introducing just the LKR gene and the DHDPS gene into a plant."

3. It was stated in paragraph 10 of my declaration previously submitted on August 24, 2000 that a co-transformation experiment in which a chimeric gene designed for co-suppression of LKR was combined with a chimeric gene for expression of lysine insensitive DHDPS was in progress. That experiment was expected to yield transformants that produced seeds with higher free lysine levels than transformants from a parallel experiment using the DHDPS gene alone. The results of those experiments have now been obtained and they do confirm the prediction that transformants comprising the chimeric gene designed for co-suppression of LKR and the chimeric gene for expression of lysine insensitive DHDPS produced seeds with higher free lysine levels than transformants from a parallel experiment using the DHDPS gene alone. These results are depicted in Figure 2 and Table 1.

4. The chimeric genes used for the experiments were:

i) corn globulin1 promoter/corn chloroplast transit sequence/ Corynebacterium dapA gene/corn globulin1 3'UTR; and

ii) corn 27kd zein promoter/fragment of corn LKR-SDH cDNA/corn 10kd zein 3' UTR

Seeds from many transformation events from each experiment were analyzed for free lysine content. It is clear from the data presented in Figure 2 that the best seeds obtained from the co-transformation experiment had considerably higher free lysine levels than the best seeds obtained from the transformation experiment where only the DHDPS gene was used. The average free lysine level from the 30 highest lysine seeds, or from the 70 highest lysine seeds, was about 2-fold higher for the co-transformation experiments compared the DHDPS only experiment.

5. It also was stated in paragraph 10 of my previous declaration submitted on August 24, 2000 that an LKR co-suppression transformant which showed an increased seed free lysine phenotype for several generations, and behaved genetically as a single locus transgene insertion, was crossed to a transgenic line that accumulates excess free lysine due to expression of lysine insensitive DHDPS and AK. Results from that experiment, which were not available at the time of the previous declaration, have confirmed the expectations expressed there, namely that seeds carrying both transgene loci will have higher free lysine levels than either parent. The data are presented in Figure 1.

6. In this experiment described in paragraph 5 above, transgenic lines homozygous for an insertion of DHDPS and AK genes, or homozygous for the co-suppressing LKR/SDH gene, were each crossed to a wild type corn line or to each

other. The F1 progeny seed from these crosses are hemizygous for the DHDPS and AK transgenic insertion, the co-suppressing LKR/SDH transgenic insertion, or both. Each cross was repeated at least 5 times, and seeds from the resulting corn ears were harvested and analyzed for free lysine levels. The results depicted in Figure 1 are averages derived from these repetitions. These results show the dramatic increase in free lysine resulting from the combination of increasing the synthesis of lysine via expression of the DHDPS gene and blocking the major pathway for lysine catabolism by co-suppressing the LKR/SDH gene.

7. Parenthetically, it is noted that a concern was raised in the Office Action dated November 22, 2000 that results from combining the DHDPS and AK transgenic insertions with a co-suppressing LKR/SDH transgenic insertion would not be predictive of combining a DHDPS only transgenic insertion with a co-suppressing LKR/SDH transgenic insertion. It is noted that there is evidence in the subject application that AK plays a secondary role to DHDPS for increasing the synthesis of lysine.

For example, it was demonstrated for (i) rapeseed transformants on page 31 at lines 18 – 24 of the specification that :

"Transformants expressing DHDPS protein showed a greater than 100-fold increase in free lysine level in their seeds. There was a good correlation between transformants expressing higher levels of DHDPS protein and those having higher levels of free lysine. One transformant that expressed AKIII-M4 in the absence of *Corynebacteria* DHDPS showed a 5-fold increase in the level of free threonine in the seeds. Concomitant expression of both enzymes resulted in accumulation of high levels of free lysine, but not threonine."

And for (ii) corn transformants (page 33 at lines 15 – 24:

"Free lysine levels in the seeds is increased from about 1.4% of free amino acids in control seeds to 15-27% in seeds of transformants expressing *Corynebacterium* DHDPS alone from the globulin 1 promoter. The increased free lysine was localized to the embryo in seeds expressing *Corynebacterium* DHDPS from the globulin 1 promoter.

The large increases in free lysine result in significant increases in the total seed lysine content. Total lysine levels can be increased at least 130% in seeds expressing *Corynebacterium* DHDPS from the globulin 1 promoter. . . . Greater increases in free lysine levels can be achieved by expressing *E. coli* AKIII-M4 protein from the globulin 1 promoter in concert with *Corynebacterium* DHDPS."

8. Thus, the gene encoding lysine insensitive AK can enhance the effect of the DHDPS gene on lysine synthesis by increasing overall flux through the biosynthetic pathway. However, AK does not increase lysine when expressed without DHDPS. It is the DHDPS gene that is necessary for increasing the synthesis of lysine. The presence of the AK gene along with the DHDPS gene in the cross described above is inconsequential with respect to proof of the concept that the **combination** of increasing lysine synthesis (which can be achieved using the DHDPS gene alone or in combination with the AK gene) and blocking lysine catabolism (which can be achieved by blocking expression of the LKR/SDH gene via co-suppression) works better than either alone.

9. The genetic cross experiment and the co-transformation experiment described above, taken together with the detailed description of the invention provided in the patent application and the previous declaration, clearly demonstrate that an increased lysine content is achieved when a lysine insensitive DHDPS gene (with or without a lysine insensitive AK gene) is combined with a co-suppressing LKR gene.

I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements are made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Saverio Carl Falco

Feb 16, 2001
Date

Figure 1: Compare DHDPS + AK, csLKR, DHDPS + AK + csLKR

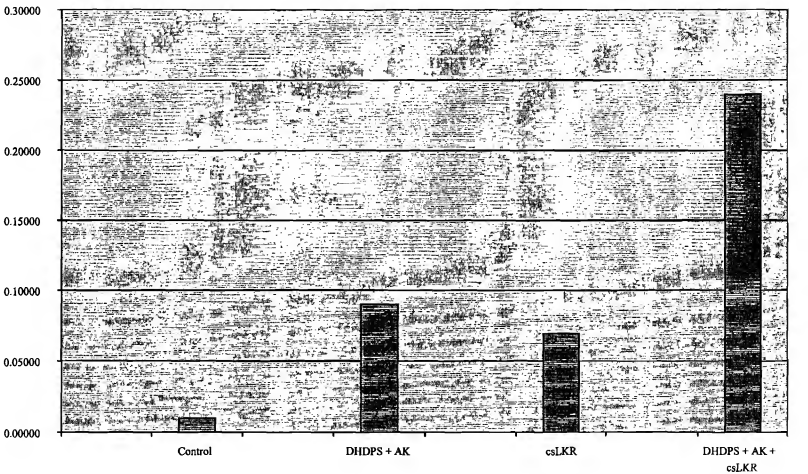


Figure 2: Comparison of seeds from transformation of DHDPS alone vs DHDPS + csLKR

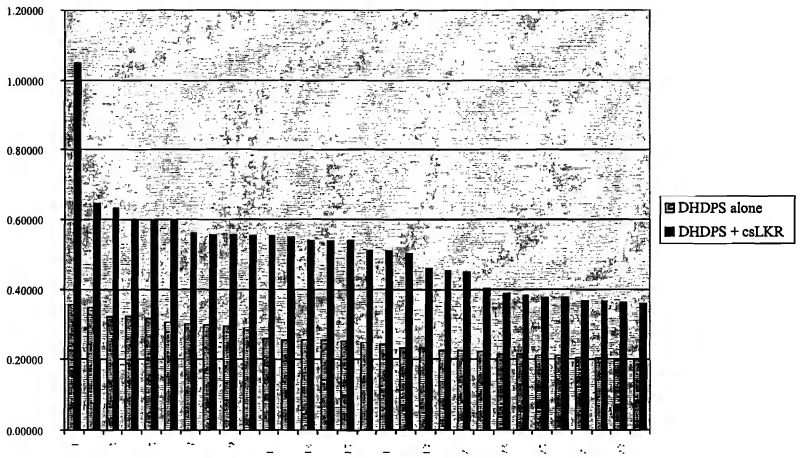


Table 1

	DHDPS alone wt % Free Lys	DHDPS + csLKR wt % Free Lys	wild type corn wt % Free Lys
Avg of best 30 seeds	**0.26**	**0.51**	**0.01**
Avg of best 70 seeds	**0.20**	**0.39**	**0.01**

Evidence Appendix D

Tang et al., Plant Cell 9:1305-1316, 1997

This reference was entered into the record by the Examiner in the office Action mailed September 28, 2007, initialed PTO form 1449.

The Plant Cell, Vol. 9, 1305–1316, August 1997

Regulation of Lysine Catabolism through Lysine–Ketoglutarate Reductase and Saccharopine Dehydrogenase in Arabidopsis

Guiliang Tang,[1] Daphna Miron,[1] Judith X. Zhu-Shimoni, and Gad Galili[2]

Department of Plant Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

In plant and mammalian cells, excess lysine is catabolized by a pathway that is initiated by two enzymes, namely, lysine–ketoglutarate reductase and saccharopine dehydrogenase. In this study, we report the cloning of an Arabidopsis cDNA encoding a bifunctional polypeptide that contains both of these enzyme activities linked to each other. RNA gel blot analysis identified two mRNA bands—a large mRNA containing both lysine–ketoglutarate reductase and saccharopine dehydrogenase sequences and a smaller mRNA containing only the saccharopine dehydrogenase sequence. However, DNA gel blot hybridization using either the lysine–ketoglutarate reductase or the saccharopine dehydrogenase cDNA sequence as a probe suggested that the two mRNA populations apparently are encoded by the same gene. To test whether these two mRNAs are functional, protein extracts from Arabidopsis cells were fractionated by anion exchange chromatography. This fractionation revealed two separate peaks—one containing both coeluted lysine–ketoglutarate reductase and saccharopine dehydrogenase activities and the second containing only saccharopine dehydrogenase activity. RNA gel blot analysis and in situ hybridization showed that the gene encoding lysine–ketoglutarate reductase and saccharopine dehydrogenase is significantly upregulated in floral organs and in embryonic tissues of developing seeds. Our results suggest that lysine catabolism is subject to complex developmental and physiological regulation, which may operate at gene expression as well as post-translational levels.

INTRODUCTION

In the cell, the level of the essential amino acid lysine is subject to tight regulation in both mammals and plants. In both types of organisms, excess lysine is catabolized via saccharopine and α-aminoadipic semialdehyde into α-aminoadipic acid and glutamate (Moller, 1976; Bryan, 1980; Markovitz et al., 1984; Galili et al., 1994; Galili, 1995; Goncalves-Butruille et al., 1996). The first enzyme in the lysine catabolic pathway is lysine–ketoglutarate reductase (LKR), which condenses lysine and α-ketoglutarate into saccharopine and uses the cofactor NADPH (Figure 1, reaction 1). The second enzyme, saccharopine dehydrogenase (SDH), converts saccharopine into α-aminoadipic semialdehyde and glutamate (Figure 1, reaction 2). This enzyme uses NAD^+ or, much less efficiently, $NADP^+$ as a cofactor (Markovitz et al., 1984; Goncalves-Butrullie et al., 1996).

The molecular and biochemical regulation of lysine catabolism is still not clearly understood. Feeding lysine to rats or applying it to tobacco plants stimulated the activity of LKR in rat livers or in tobacco seeds, respectively (Foster et al., 1993; Karchi et al., 1994). Stimulation of this enzyme has also been observed in transgenic tobacco seeds overproducing lysine because of expression of a feedback-insensitive bacterial dihydrodipicolinate synthase (Karchi et al., 1995). This suggests that in both mammalian and plant cells, lysine may autoregulate its own catabolism. In addition, recent studies have shown that in tobacco seeds, the lysine-dependent stimulation of LKR activity is mediated by an intracellular signaling cascade requiring Ca^{2+} and protein phosphorylation (Karchi et al., 1995). The control of LKR activity in plants may be even more complex. In developing maize seeds, LKR activity was found to be reduced by two- to threefold in the high-lysine *opaque2* mutant, as compared with wild-type plants (Brochetto-Braga et al., 1992). *Opaque 2* is a transcription factor that regulates the expression of seed storage proteins (Shotwell and Larkins, 1988). This transcription factor could also complement the yeast GCN4 transcription factor that regulates the expression of many yeast genes encoding enzymes involved in amino acid metabolism (Hinnebusch, 1988).

Although LKR and SDH appear to control important processes, their structural aspects and cellular functions differ among various eukaryotic species. In yeast cells, in which lysine is synthesized via α-aminoadipate (Bhattacharjee, 1985), LKR and SDH play essential roles in lysine biosynthesis, and they appear as two separate polypeptides (Ogawa and Fujioka, 1978). In mammalian cells, which cannot synthesize lysine, LKR (LYS1) and SDH (LYS9) play an essential role in the catabolism of excess cellular lysine (Dancis et al., 1969), but their structural aspects may vary among species.

[1] Both authors contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail lpgad@wicc mail.weizmann.ac.il; fax 972-8-9344181.

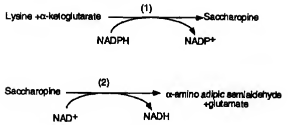


Figure 1. Lysine Catabolism via the Saccharopine Pathway.

Reaction 1 is catalyzed by LKR, which condenses L-lysine and α-ketoglutarate into saccharopine. Reaction 2 is catalyzed by SDH, which hydrolyzes saccharopine into α-aminoadipic semialdehyde and glutamic acid.

In rat liver, LKR and SDH were shown to be distinct monofunctional enzymes (Noda and Ichihara, 1978); however, human placenta (Fjellstedt and Robinson, 1975) and bovine liver (Markovitz et al., 1984) possess these two enzyme activities on a single bifunctional protein. In plants, which synthesize lysine via diaminopimelate, LKR and SDH also function in lysine catabolism (Arruda and da Silva, 1983), and recently, a bifunctional LKR/SDH enzyme has been purified from developing maize seeds (Brochetto-Braga et al., 1992). Moreover, in plants, LKR and SDH activities have been detected only in developing seeds to date (Arruda and da Silva, 1983; Karchi et al., 1994).

To elucidate further the regulatory role of LKR and SDH in lysine catabolism, we have cloned and characterized two cDNAs encoding a bifunctional LKR/SDH and a monofunctional SDH from Arabidopsis. We also show that Arabidopsis cells contain an mRNA species encoding a bifunctional LKR/SDH and another mRNA encoding a monofunctional SDH and that these are likely to be transcribed from a single gene. In addition, we have determined that expression of the Arabidopsis LKR/SDH gene is subject to spatial and developmental controls.

RESULTS

Identification and Characterization of an Arabidopsis cDNA Encoding a Monofunctional SDH

From the Arabidopsis sequence databases, we have identified an expressed sequence tag (EST) clone showing significant homology with SDH from yeast (clone 23A3T7). The complete nucleotide sequence of this cDNA was determined. As shown in Figure 2, this ~1.5-kb cDNA (designated cAt-SDH) contains an open reading frame initiated by an ATG codon with a consensus (G/A) at position −3 (Joshi, 1987) and encodes a putative protein of 482 amino acids. In addition, this cDNA contains a 5′ noncoding sequence of 51 nucleotides and a 3′ noncoding sequence of 80 nucleotides that is ended by a poly(A) tail. As shown in Figure 3, the open reading frame of cAt-SDH has significant homology with yeast SDH (LYS9), sharing 36.1% identity and 56.4% similarity. The initiation ATG and the stop codons of both yeast and the putative Arabidopsis SDH also appeared at

```
   1 CTGAAGATTGGTAAAGTCCAGCAAGAAAATGAGATAAAAGAGAAGCCTGAAATGACGAAA
                                                        M  T  K

  61 AAATCAGGTGTTTTGATTCTTGGTGCTGGACGTGTGTGTCGCCCAGCTGCTGATTTCCTA
      K  S  G  V  L  I  L  G  A  G  R  V  C  R  P  A  A  D  F  L

 121 GCTTCAGTTAGAACCATTTCGTCACAGCAATGGTACAAAACATATTTCGGAGCAGACTCT
      A  S  V  R  T  I  S  S  Q  Q  W  Y  K  T  Y  F  G  A  D  S

 181 GAAGAGAAAACAGATGTTCATGTGATTGTCGCGTCTCTGTATCTTAAGGATGCCAAAGAG
      E  E  K  T  D  V  H  V  I  V  A  S  L  Y  L  K  D  A  K  E

 241 ACGGTTGAAGGTATTTCAGATGTAGAAGCAGTTCGGCTAGATGTATCTGATAGTGAAAGT
      T  V  E  G  I  S  D  V  E  A  V  R  L  D  V  S  D  S  E  S

 301 CTCCTTAAGTATGTTTCTCAGGTTGATGTTGTCCTAAGTTTATTACCTGCAAGTTGTCAT
      L  L  K  Y  V  S  Q  V  D  V  V  L  S  L  L  P  A  S  C  H

 361 GCTGTTGTAGCAAAGACATGCATTGAGCTGAAGAAGCATCTCGTCACTGCTAGCTATGTT
      A  V  V  A  K  T  C  I  E  L  K  K  H  L  V  T  A  S  Y  V

 421 GATGATGAAACGTCCATGTTACATGAGAACGCTAAGAGTGCTGGGATAACGATTCTAGGC
      D  D  E  T  S  M  L  H  E  N  A  K  S  A  G  I  T  I  L  G

 481 GAAATGGGACTGGACCCTGGAATCGATCCCATGATGGCGAAGAAAATGATCAACGATGCT
      E  M  G  L  D  P  G  I  D  P  M  M  A  K  K  M  I  N  D  A

 541 CCTATCACAAAAGGGAAAGTGAAGTCTTTTACCTCTTATTGTGGAGGGCTTCCCTCTCGT
      P  I  T  K  G  K  V  K  S  F  T  S  Y  C  G  G  L  P  S  R

 601 GCTGCAGCAAATAATCCATTAGCATATAAATTTAGCTGGAACCCTGCTGGAGCAATTCGA
      A  A  A  N  N  P  L  A  Y  K  F  S  W  N  P  A  G  A  I  R

 661 GCTGGTCAAAACCCCGCCAAATACAAAAGCAACGGCGACATAATACATGTTGATGGGAAG
      A  G  Q  N  P  A  K  Y  K  S  N  G  D  I  I  H  V  D  G  K

 721 AATCTCTATGATTCCGCGGCAAGATTCCGAGTACCTAATCTTCCAGCTTTTGCATTGGAG
      N  L  Y  D  S  A  A  R  F  R  V  P  N  L  P  A  F  A  L  E

 781 TGTTTTCCAAATCGTGACTCCTTGGTTTACGGGGAACATTATGGCATCGAGAGCGAAGCA
      C  F  P  N  R  D  S  L  V  Y  G  E  H  Y  G  I  E  S  E  A

 841 ACAACGATATTTCGTGGAACACTCAGATATGAAGGGTTTAGTATGATAATGGCAACACTT
      T  T  I  F  R  G  T  L  R  Y  E  G  F  S  M  I  M  A  T  L

 901 TCGAAACTTGGATTCTTTGACAGTGAAGCAAATCAAGTACTCTCCACTGGAAAGAGGATT
      S  K  L  G  F  F  D  S  E  A  N  Q  V  L  S  T  G  K  R  I

 961 ACGTTTGGTGCTCTTTTAAGTAACATTCTAAATAAGGATGCAGACAATGAATCAGAGCCC
      T  F  G  A  L  L  S  N  I  L  N  K  D  A  D  N  E  S  E  P

1021 CTAGCGGGAGAAGAAGAGATAAGCAAGAGAATTATCAAGCTTGGACATTCCAAGGAGACT
      L  A  G  E  E  E  I  S  K  R  I  I  K  L  G  H  S  K  E  T

1081 GCAGCCAAAGCTGCCAAAACAATTGTATTCTTGGGGTTCAACGAAGAGAGGGAGGTTCCA
      A  A  K  A  A  K  T  I  V  F  L  G  F  N  E  E  R  E  V  P

1141 TCACTGTGTAAAAGCGTATTTGATGCAACTTGTTACCTAATGGAAGAGAAACTAGCTTAT
      S  L  C  K  S  V  F  D  A  T  C  Y  L  M  E  E  K  L  A  Y

1201 TCCGGAAATGAACAGGACATGGTGCTTTTGCATCACGAAGTAGAAGTGGAATTCCTTGAA
      S  G  N  E  Q  D  M  V  L  L  H  H  E  V  E  V  E  F  L  E

1261 AGCAAACGTATAGAGAAGCACACTGCGACTCTTTTGGAATTCGGGGACATCAAGAATGGA
      S  K  R  I  E  K  H  T  A  T  L  L  E  F  G  D  I  K  N  G

1321 CAAACAACAACCGCTATGGCCAAGACTGTTGGGATCCCTGCAGCCATTGGAGCTCTGCTG
      Q  T  T  T  A  M  A  K  T  V  G  I  P  A  A  I  G  A  L  L

1381 TTAATTGAAGACAAGATCAAGACAAGAGGAGTCTTAAGGCCTCTCGAAGCAGAGGTGTAT
      L  I  E  D  K  I  K  T  R  G  V  L  R  P  L  E  A  E  V  Y

1441 TTGCCAGCTTTGGATATATTGCAAGCATATGGTATAAAGCTGATGGAGAAGGCAGAATGA
      L  P  A  L  D  I  L  Q  A  Y  G  I  K  L  M  E  K  A  E  *

1501 TCAAAGAACTCTGTATATTGTTTCTCTCTATAACTTGGAGTTGGAGACAAAGCTGAAGAA

1561 GACAGAGACATTAGACCAGCAAAAAAAAAAAAAAAA
```

Figure 2. Nucleotide and Deduced Amino Acid Sequence of cAt-SDH.

The ATG and TAG initiation and stop codons of the open reading frame encoding the putative SDH protein are in boldface and underlined. The asterisk indicates the protein termination site. The GenBank accession number is U90523.

```
  1 MTKKSGVLILGAGRVCRPAADFLASVRTISSQQWYKTYFGADSEEKTDVH  50
       . ..||:||.| |..|. | ||.                    ..|::
  1 ..MGKNVLLLGSGFVAQPVIDTLAA....................NDDIN  28

 51 VIVASLYLKDAKETVEGISDVEAVRLDVSDSESLLKYVSQVDVVLSLLPA 100
    |.||:   | :| :.:.  |: .|:.|||.|...| | :.: |||:||:|
 29 VTVACRTLANA.QALAKPSGSKAISLDVTDDSALDKVLADNDVVISLIPY  77

101 SCHAVVAKTCIELKKHLVTASYVDDETSMLHENAKSAGITILGEMGLDPG 150
     . |: |.|..|  |..:||.||:.... . |..:  .||||::.|:|||||
 78 TFHPNVVKSAIRTKTDVVTSSYISPALRELEPEIVKAGITVMNEIGLDPG 127

151 IDPMMAKKMINDAPITKGKVKSFTSYCGGLPSRAAANNPLAYKFSWNPAG 200
    ||.: | | |::.. . ||:||| |||||||.....:|||:|||||.. |
128 IDHLYAVKTIDEVHRAGGKLKSFLSYCGGLPAPEDSDNPLGYKFSWSSRG 177

201 AIRAGQNPAKYKSNGDIIHVDGKNLYDSAARFRVPNLPAFALECPPNRDS 250
    .: | .|.|||..:|.| |.:.:| ..| .: :  .|::|: |:||||||
178 VLLALRNSAKYWKDGKIETVSSEDLMATAKPYFI..YPGYAFVCYPNRDS 225

251 LVYGEHYGIESEATTIFRGTLRYEGFSMIMATLSKLGFFDSEANQVLSTG 300
     :: : | | .||.|::|||||:||. :: .| .:|::..:||:::|
226 TLFKDLYHI.PEAETVIRGTLRYQGFPEFVKALVDMGMLKDDANEIFS.. 272

301 KRITFGALLSNILNKDADNESEPLAGEEEISKRIIKLGHSKETAAKAAKT 350
    |.|.:.. |...|. ......: :|: :. ..:      ...|. .:  ..
273 KPIAWNEALKQYLGAKSTSKEDLIASIDSKATW.....KDDEDRERILSG 317

351 IVFLGFNEEREVPSLCKSVFDATCYLMEEKLAYSGNEQDMVLLHHEVEVE 400
    :.:||: .: .:.. ... :|. |  :|| :.|.:||.|||:|:|..::|
318 FAWLGLFSDAKITPRGNA.LDTLCARLEELMQYEDNERDMVVLQHKFGIE 366

401 FLESKRIEKHTATLLEFGDIKNGQTTTAMAKTVGIPAAIGALLLIEDKIK 450
    : ::. .|.:|.||:::|.:      .. ..|| |||.|.||:. :::::.||
367 WADGT.TETRTSTLVDYGKV...GGYSSMAATVGYPVAIATKFVLDGTIK 412

451 TRGVLRPLEAEVYLPALDIL.QAYGIKLMEKAE* 482
    ..|:| |..:|:  | :. | : ||| |.||.
413 GPGLLAPYSPEINDPIMKELKDKYGIYLKEKTVA* 446
```

Figure 3. Comparison of the Deduced Amino Acid Sequence of cAt-SDH with the Yeast SDH.

The top line indicates the amino acid sequence of cAt-SDH; the bottom line is that of the yeast SDH (LYS9). Identical amino acids are indicated by bars; highly similar amino acids are indicated by colons; and similar amino acids are indicated by a single dot. The asterisks indicate the protein termination sites.

very similar positions along the open reading frames. However, the Arabidopsis SDH has a short amino acid sequence (Figure 3, positions 26 to 45) that is not present in the yeast SDH. The 5′ noncoding region of cAt-SDH contains an additional ATG consensus codon in a coding frame different from that of the putative SDH open reading frame and is immediately followed by a stop codon. Whether this ATG has any functional role is still not known.

To test whether cAt-SDH encodes an SDH enzyme, the entire coding sequence of this cDNA was subcloned into either the pUC18 or pET-15b bacterial expression vectors and used to transform *Escherichia coli* cells. As shown in Figure 4, bacterial cells harboring either of these plasmids containing the cAt-SDH insert have significantly elevated levels of SDH activity, as compared with control bacteria harboring the expression plasmids with no inserts.

Analysis of SDH mRNA Levels in Different Arabidopsis Tissues

To test the expression of the Arabidopsis SDH gene in different tissues, total RNA was extracted from cell cultures, leaves, stems, roots, flowers, and young seedlings, and the levels of SDH mRNA were analyzed by hybridization with the cAt-SDH DNA as a probe. As shown in Figure 5, two major cross-hybridizing mRNA bands were detected. One had the expected size of ~1.5 kb corresponding to cAt-SDH, and a second larger mRNA was ~3.5 kb. Both mRNA bands were detected in all tissues after a long exposure time (data not shown) and were most intense in flowers.

Cloning and Characterization of an Arabidopsis cDNA Encoding a Putative Bifunctional LKR/SDH Polypeptide

Previous studies have shown that plants, like mammals, may have bifunctional LKR/SDH enzymes (Goncalves-Butruille et al., 1996). Therefore, we hypothesized that the ~3.5-kb mRNA band, shown in Figure 5, encodes a bifunctional LKR/SDH in which the SDH region is highly homologous to the monofunctional SDH mRNA. To examine this possibility, we screened an Arabidopsis cDNA library with the 1.5-kb SDH cDNA as a probe. Several positive clones were rescued, and the longest (~3.2-kb insert, designated cAt-LKR/SDH) was sequenced. As shown in Figure 6, cAt-LKR/SDH contains a long open reading frame of 3195 nucleotides encoding a protein of 117 kD. The size of the encoded protein is similar to the size of the bifunctional LKR/SDH recently purified from

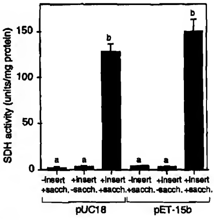


Figure 4. cAt-SDH Encodes a Functional SDH Enzyme.

The coding sequence of cAt-SDH was subcloned into two different bacterial expression vectors and transformed into *E. coli*. Protein extracts from bacteria harboring the plasmids containing cAt-SDH, as well as control bacteria transformed with the expression vector lacking cAt-SDH, were analyzed for SDH activity with (+) or without (−) the substrate saccharopine (sacch.). Each histogram represents an average of three separate activity tests ±SE. Letters above error bars represent significant differences at the 5% level, as determined by an ANOVA test. −insert, the empty vector not containing cAt-SDH; +insert, the vector with cAt-SDH.

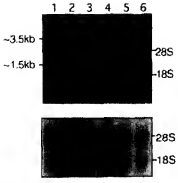


Figure 5. RNA Gel Blot Analysis of cAt-SDH mRNA.

Twenty micrograms of total RNA from cell cultures (lane 1), young seedlings (lane 2), floral organs (lane 3), leaves (lane 4), stems (lane 5), and roots (lane 6) was fractionated by gel electrophoresis and either hybridized on an RNA gel blot with cAt-SDH used as a probe (top) or stained with ethidium bromide as a control (bottom). The migration of the 18S and 28S rRNAs is shown at right. The positions of the monofunctional SDH mRNA (~1.5 kb) and the bifunctional LKR/SDH mRNA (~3.5 kb) are indicated at left.

maize (125 kD; Goncalves-Butruille et al., 1996) and from soybean (123 kD; D. Miron, S. Ben-Yaacov, D. Reches, and G. Galili, manuscript in preparation). This open reading frame is flanked by a 5′ noncoding sequence of 62 nucleotides and a 3′ noncoding sequence of 10 nucleotides. The cAt-LKR/SDH cDNA also lacks a 3′ poly(A) tail, suggesting that its 3′ region is not complete. Interestingly, the 3′ 1510 nucleotides of cAt-LKR/SDH are 100% homologous to nucleotides 1 to 1510 of cAt-SDH encoding the monofunctional SDH (cf. Figure 2).

As shown in Figure 7, the N-terminal part of the putative protein encoded by cAt-LKR/SDH (460 amino acids) exhibits significant homology to the yeast monofunctional LKR, with 24.9% identity and 52.1% similarity. The ATG initiation and stop codons of the yeast and the putative Arabidopsis LKR proteins also appear at comparable places along the open reading frame (Figure 7). However, the Arabidopsis LKR also has several small amino acid sequences that are not present in the yeast LKR (Figure 7).

The 5′ noncoding region of cAt-LKR/SDH contains three ATG triplets located seven to 41 nucleotides upstream of the presumed ATG translation initiation codon of the LKR/SDH open reading frame. These ATG codons form small open reading frames of nine to 15 amino acids, and none of these ATG codons contains the (A/G) consensus at position −3, which is generally found before eukaryotic translation initiation codons (Joshi, 1987), suggesting that these ATG triplets may have limited if any function in translational initiation.

Amino acid sequence alignment of the deduced polypeptide product of cAt-LKR/SDH with the yeast monofunctional LKR and SDH (Figures 2 and 6) shows that the putative cAt-LKR/SDH–encoded protein contains an intermediate region (amino acids 462 to 582, shown in boldface letters in Figure 6) that is not present in either the yeast LKR or the SDH enzymes. Although the functional significance of this region is still not known, intermediate regions previously have been found in other bifunctional polypeptides, such as the aspartate kinase/homoserine dehydrogenase isozyme of the aspartate family pathway (Ghislain et al., 1994; Galili, 1995).

To test whether the ~3.5-kb mRNA detected on the RNA gel blot shown in Figure 5 is related to cAt-LKR/SDH, the same blot was washed to remove the cAt-SDH probe and rehybridized with the putative LKR domain of cAt-LKR/SDH. As shown in Figure 8, this hybridization detected the ~3.5-kb mRNA band corresponding to cAt-LKR/SDH but not the ~1.5-kb mRNA band corresponding to cAt-SDH.

To determine further whether the N-terminal part of cAt-LKR/SDH encodes an LKR enzyme, the entire coding sequence of this cDNA was subcloned into the bacterial expression vector pET-15b and used to transform *E. coli* cells. Bacterial cells harboring this plasmid had SDH but no LKR activity (data not shown). Because bacterial cells did not produce an active LKR, we attempted to express the Arabidopsis LKR protein in yeast cells. Yeast has a monofunctional LKR enzyme, so we subcloned the N terminus of the presumed LKR domain of cAt-LKR/SDH into the yeast expression vector pVT-102u and transformed this plasmid into the yeast *Lys1* mutant. As shown in Figure 9, yeast cells harboring this plasmid have significantly higher LKR activity than do control cells transformed with the same plasmid without the LKR insert, thereby confirming our supposition that cAt-LKR/SDH indeed encodes a bifunctional LKR/SDH enzyme.

Organization of the LKR and SDH Genes in Arabidopsis

Based on the DNA sequence identity between cAt-SDH and the 3′ half of cAt-LKR/SDH (cf. Figures 2 and 6) and the presence of two mRNA species, corresponding in sizes to both cAt-SDH and cAt-LKR/SDH (Figure 5), we wanted to determine whether these two cDNAs are clustered within a single locus. To investigate whether the two cDNAs were derived from a single gene, Arabidopsis DNA was digested with several restriction enzymes, fractionated by agarose gel electrophoresis, and hybridized on DNA gel blots by using the SDH cDNA as a probe. After 1 week of autoradiography, the membrane was stripped and rehybridized with the LKR probe. As illustrated in Figures 10A and 10B, a comparison of the two autoradiographies shows that a signal appeared at exactly the same position when digested with both EcoRI and BamHI. These results suggest that the cAt-SDH/LKR and cAt-SDH are derived from a single gene. In the HindIII and BglII digests, the LKR and SDH probes highlighted different bands, apparently because the Arabidopsis LKR/SDH gene has a number of introns (G. Tang and G. Galili, unpublished data) that contain single or multiple sites for some of the restriction enzymes used for digestions.

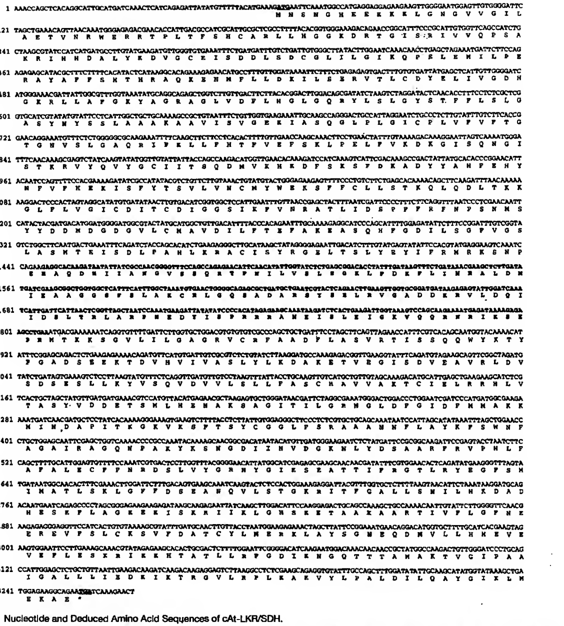


Figure 6. Nucleotide and Deduced Amino Acid Sequences of cAt-LKR/SDH.

The ATG and TAG translation initiation and stop codons of the open reading frame encoding the putative LKR/SDH protein are in boldface and underlined. The boldface region in the middle of the cDNA represents an intermediate region not present in either of the monofunctional LKR and SDH proteins of yeast. The asterisk indicates the protein termination site. The GenBank accession number is U90522.

Primer Extension Analysis of the Monofunctional SDH mRNA

To characterize further the putative monofunctional SDH mRNA observed on the RNA gel blots, we synthesized a 26-bp antisense DNA primer homologous to a region located 20 nucleotides downstream of the ATG translation initiation codon of cAt-SDH (Figure 2, nucleotides 75 to 100). This primer was then hybridized with total RNA from Arabidopsis flowers, and the hybrid molecules were used as templates for

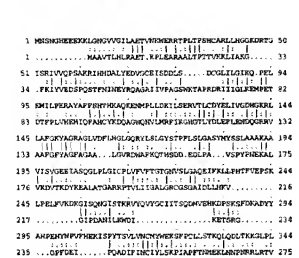


Figure 7. Comparison of the Deduced Amino Acid Sequence of cAt-LKR/SDH with the Yeast LKR.

The top line indicates the amino acid sequence of the LKR domain of the cAt-LKR/SDH; the bottom line is that of the yeast LKR (LYS1). Identical amino acids are indicated by bars; highly similar amino acids are indicated by colons; and similar amino acids are indicated by a single dot. The asterisk indicates the protein termination site.

reverse transcription in a primer extension reaction. As shown in Figure 11, this reaction generated a DNA band of 54 nucleotides that was extended approximately five nucleotides upstream of the cAt-SDH translation initiation ATG codon.

Arabidopsis Cells Contain Bifunctional LKR/SDH and Monofunctional SDH Isozymes

To determine whether the two mRNAs derived from the Arabidopsis LKR/SDH gene were functional in translating bifunctional LKR/SDH and monofunctional SDH isozymes, we partially purified LKR and SDH from an Arabidopsis cell culture by using an anion exchange column, after polyethylene glycol (PEG) fractionation. As shown in Figure 12, elution from the anion exchange column resolved two distinct SDH peaks. The first was eluted at ~90 mM KCl and contained only SDH activity, whereas the second peak was eluted at ~190 mM KCl and had both SDH and LKR activities. The level of SDH activity in the peak that did not show LKR activity was ~3.5-fold higher than the level in the peak containing both coeluted SDH and LKR activities. Moreover, under the excess substrate concentrations that were used in the enzymatic assays (D. Miron, S. Ben-Yaacov, D. Reches, and G. Galili, manuscript in preparation), LKR activity in this peak was approximately fourfold higher than was SDH activity.

In Situ Hybridization with the SDH and LKR mRNAs as Probes

We have shown that cAt-SDH mRNA is expressed to a high level in floral tissues of Arabidopsis (Figure 5). To determine whether the expression of both LKR/SDH and SDH mRNAs in Arabidopsis tissues is subject to developmental regulation, particularly in reproductive organs, we used LKR/SDH RNA probes for in situ hybridization analysis of Arabidopsis flowers and seeds. Digoxigenin-labeled RNA probes from both LKR (Figures 13A, 13D, and 13G) and SDH (Figures 13B, 13E, and 13H) domains of the Arabidopsis LKR/SDH cDNA were used in this analysis. As shown in Figures 13A and 13B, the LKR and SDH mRNA was highly abundant in the ovules and vascular tissue of anther filaments but not in pollen grains. In developing and mature seeds, hybridization signals were found in the embryo (at either the globular [Figures 13G and 13H] or torpedo [Figures 13D and 13E] stages) and in the outer layers of the endosperm (Figures 13G and 13H). No signal was detected in the control sections reacted with either the LKR (Figure 13C) or SDH (Figure 13F) sense probes. The somewhat lower intensity of signal obtained with the SDH probe compared with that of the LKR probe was probably due to a lower amount of the SDH probe and possibly the lower incorporation of digoxigenin dunng in vitro transcription used during hybridization. This result indicates that the expression of both SDH and LKR/SDH genes is regulated in a tissue-specific manner during plant development.

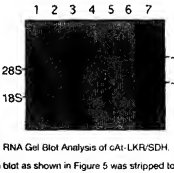


Figure 8. RNA Gel Blot Analysis of cAt-LKR/SDH.

The same blot as shown in Figure 5 was stripped to remove the cAt-SDH probe and rehybridized, with the LKR coding region of cAt-LKR/SDH as a probe. Lane 1 contains RNA from cell cultures; lane 2, young seedlings; lane 3, floral organs; lane 4, leaves; lane 5, stems; and lane 6, roots. Lane 7 is the same as lane 3 shown in Figure 5 containing RNA from floral organs hybridized with cAt-SDH as a probe. The migration of the 18S and 28S rRNAs is shown at left. The positions of the monofunctional SDH mRNA (~1.5 kb) and the bifunctional LKR/SDH mRNA (~3.5 kb) are shown at right.

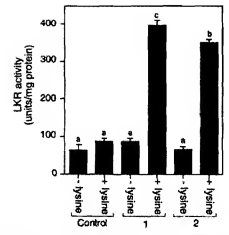


Figure 9. The LKR Domain of cAt-LKR/SDH Encodes a Functional LKR Enzyme in Yeast Cells.

The putative LKR domain of cAt-SDH was subcloned into a yeast expression vector and transformed into yeast. Protein extracts from two different yeast colonies (marked 1 and 2) harboring the plasmid containing the Arabidopsis LKR, as well as control yeast cells transformed with the expression vector without the insert, were then analyzed for LKR activity in reactions containing (+lysine) or lacking (−lysine) the substrate lysine. Letters above error bars represent significant differences at the 5% level, as determined by an ANOVA test. Each histogram is an average of three separate activity determinations ±SE.

DISCUSSION

Arabidopsis Contains Bifunctional LKR/SDH and Monofunctional SDH Isozymes, Which May Be Derived from a Single Gene

This report describes the cloning of LKR and SDH cDNAs from Arabidopsis and shows that the structural and regulatory aspects of LKR and SDH in plants are much more complex than what has been previously elucidated for yeast and mammals (Bhattacharjee, 1965; Feller et al., 1994). To date, either single LKR or SDH (yeast and rat) or bifunctional LKR/SDH (human, bovine, maize, and soybean) has been shown to exist within a given species; however, in this study, we show that Arabidopsis cells contain two isozymic peaks, as deduced from anion exchange chromatography. One of these peaks contains both LKR and SDH activities, which presumably are located on a bifunctional polypeptide encoded by cAt-LKR/SDH, and the other contains only SDH activity. Although a bifunctional LKR/SDH enzyme has been reported previously in maize, our results show that plant cells may also contain a monofunctional SDH. In fact, we have recently purified the SDH protein (shown in Figure 12 as the first SDH activity peak) to homogeneity and found that it is a 53-kD protein, in agreement with the expected size of a monofunctional SDH (data not shown).

Our results also strongly suggest that these two isozymes of LKR/SDH and monofunctional SDH are translated from two distinct mRNAs, which are produced from a single gene. We reached this conclusion based on several lines of evidence: (1) detection of two mRNA bands with the expected sizes of the isozymes (~1.5 and ~3.5 kb) on RNA gel blots hybridized with the monofunctional SDH cDNA as a probe under high-stringency conditions; (2) the presence of an in-frame "plant" ATG consensus codon at the initiation of the SDH coding sequence (as deduced from amino acid sequence homology with the yeast SDH), which also gave rise to the production of an active recombinant monofunctional SDH in bacteria; and (3) DNA gel blot analysis, which suggested the presence of only a single gene in Arabidopsis that hybridized with either the LKR or the SDH domains of cAt-LKR/SDH as probes.

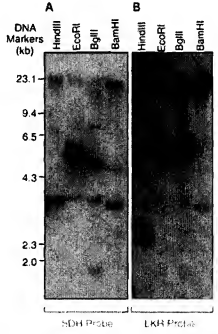


Figure 10. DNA Gel Blot Hybridization Pattern of cAt-SDH and cAt-LKR/SDH.

(A) Genomic DNA was digested with several restriction enzymes as indicated above the gel. Ten micrograms of digested DNA was separated on a gel, transferred to a membrane, and hybridized under high-stringency conditions with cAt-SDH.

(B) The same blot as shown in **(A)** was stripped and hybridized under high-stringency conditions with the LKR domain of cAt-LKR/SDH as probes.

The migration of the molecular length markers is indicated at left, and their lengths are given in kilobases.

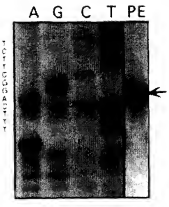


Figure 11. Primer Extension Reaction of Total Arabidopsis Flower RNA with the Antisense Primer Located 20 to 46 Nucleotides Downstream of the cAt-SDH ATG Initiation Codon.

The primer extension (PE) reaction product is indicated by an arrow; A, G, C, and T indicate sequencing ladders of the same primer annealed to the relevant genomic fragment. The sequence around the extended product is indicated at left.

The presence of an mRNA encoding a monofunctional SDH was also supported by the primer extension analysis shown in Figure 11. However, the primer extension band was shorter than expected, based on the 5′ noncoding sequence of cAt-SDH, and terminated approximately five nucleotides upstream of the translation initiation ATG of this cDNA. The reason for the shorter than expected primer extension fragment is still not known. However, computer analysis predicted that the 5′ noncoding region of cAt-SDH may contain a relatively stable stem and loop structure (data not shown). Experiments are now in progress in our laboratory to analyze whether this region may indeed form stable secondary structures in vivo and whether these structures may function in the regulation of the LKR/SDH gene expression. Nevertheless, based on the primer extension results, we cannot yet affirm whether cAt-SDH was derived from the monofunctional SDH mRNA or is a truncated form of cAt-LKR/SDH.

Structural and Functional Properties of the Bifunctional LKR/SDH Enzyme

Amino acid sequence alignment of cAt-LKR/SDH with the yeast monofunctional LKR and SDH isozymes revealed that the plant bifunctional enzyme possesses an intermediate region between the two enzyme domains that was not present in any of the yeast enzymes. Similar intermediate regions were also reported for other bifunctional enzymes, such as bacterial and plant aspartate kinase/homoserine dehydrogenase (Kalinowski et al., 1991; Ghislain et al., 1994). The functional role of this intermediate region is still not known. However, the fact that the LKR and SDH domains of the bifunctional LKR/SDH can be dissected into single functional enzymes (Figures 4 and 9; Markovitz and Chuang, 1987; Goncalves-Butruille et al., 1996) suggests that this region may enable independent folding of the two domains. In addition, because bifunctional LKR/SDH are generally homooligomers (Markovitz et al., 1984; Goncalves-Butruille et al., 1996), the intermediate domain may also function in its assembly, as was previously reported for the bacterial bifunctional aspartate kinase/homoserine dehydrogenase enzyme (Kalinowski et al., 1991).

Another interesting issue is whether the linkage between the LKR and SDH domains has a regulatory significance, which may result from "cross-talk" between the two domains. Although this issue is still not solved, our study indicates that such cross-talk may indeed occur. Upon fractionation on the anion exchange column and analysis under conditions of excess substrates of LKR and SDH (D. Miron, S. Ben-Yaacov, D. Reches, and G. Galili, manuscript in preparation), the specific activity of SDH in the monofunctional SDH peak was much higher than that in the bifunctional LKR/SDH peak. This difference could not be explained by the differential degree of purification of the two peaks because both peaks contained comparable levels of total protein. The differences in SDH activity between the two isozymes also could not be explained by differences in mRNA levels because the intensity of the LKR/SDH mRNA band was slightly higher than that of the monofunctional SDH mRNA (Figure 5). Thus, although we cannot yet rule out the possibility of variation in translational efficiency or protein stability, it is tempting to hypothesize that the activity of SDH may be negatively reg-

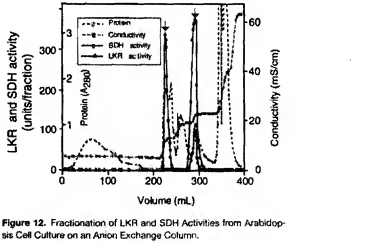


Figure 12. Fractionation of LKR and SDH Activities from Arabidopsis Cell Culture on an Anion Exchange Column.

PEG-fractionated Arabidopsis cell culture extract was loaded onto a DEAE–Sepharose column, washed, and eluted with a step gradient of 0 to 1 M KCl. The protein level, conductivity, and LKR and SDH activities in each fraction are presented. mS, millisiemens.

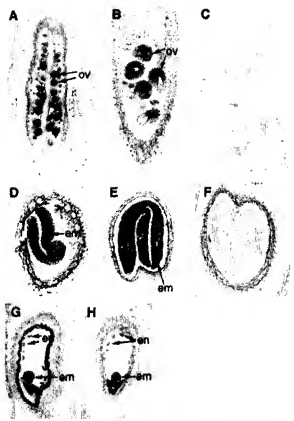


Figure 13. In Situ Hybridization of Arabidopsis Flower and Seed Tissues with LKR and SDH Antisense Probes.

(A), **(D)**, and **(G)** LKR probe. **(A)** shows a longitudinal section of an Arabidopsis flower. **(D)** shows cross-sections of seeds with a torpedo-shaped embryo. **(G)** depicts cross-sections of seeds with a globular-shaped embryo.

(B), **(E)**, and **(H)** SDH probe. **(B)** shows a longitudinal section of an Arabidopsis flower. **(E)** shows cross-sections of seeds with a torpedo-shaped embryo. **(H)** shows cross-sections of seeds with a globular-shaped embryo.

(C) and **(F)** Negative controls with LKR and SDH sense probes, respectively. **(C)** shows a longitudinal section of an Arabidopsis flower. **(F)** shows cross-sections of seeds with a torpedo-shaped embryo.

em, embryo; en, endosperm; ov, ovules.

ulated by its linked LKR domain. If indeed such a control occurs in vivo, it is expected that plant species producing only a single bifunctional LKR/SDH will accumulate saccharopine (the product of LKR and the substrate of SDH; see Figure 1), whereas those producing both isozymes will accumulate a downstream metabolite of the catabolic pathway. Interestingly, whereas lysine-overproducing transgenic soybean seeds, expressing a bacterial dihydrodipicolinate synthase, were shown to accumulate saccharopine, transgenic tobacco and canola expressing the same bacterial enzyme accumulated the downstream metabolite α-aminoadipic acid (Falco et al., 1995). Whether the differential accumulation of saccharopine and α-aminoadipic acid in these plant species is related to differential expression of the LKR/SDH and SDH isozymes still remains to be demonstrated.

Expression of the LKR/SDH Gene Is Developmentally Regulated

Although the LKR/SDH and monofunctional SDH mRNAs were detected in all tissues tested, their levels varied among the different tissues. Both mRNAs were significantly higher in floral organs than in vegetative tissues (Figure 5). In addition, in situ mRNA hybridization using reproductive organs showed that these mRNAs were most abundant in the ovaries of developing flowers as well as in the embryos but not in the endosperm tissues of developing and mature seeds. The spatial pattern of LKR/SDH gene expression in developing flowers and seeds appears very similar to that of the Arabidopsis gene encoding the bifunctional aspartate kinase/homoserine dehydrogenase that leads to the synthesis of lysine as well as threonine, methionine, and isoleucine (Zhu-Shimoni et al., 1997). These results support our previous hypothesis (Karchi et al., 1994) that expression of genes encoding enzymes in lysine biosynthesis and catabolism may be coordinately expressed during plant development. We have also previously shown that the presence of excess cellular lysine caused the stimulation of LKR activity in developing tobacco seeds (Karchi et al., 1995). Therefore, it will be interesting to test whether the coordinated expression of the LKR/SDH gene with other genes encoding enzymes in lysine biosynthesis is due to common transcriptional elements in their promoters or to a special regulation of LKR/SDH gene expression by sensing the relatively high lysine levels in cells in which lysine biosynthesis is upregulated.

Post-Transcriptional Regulation of LKR

The Arabidopsis SDH was active when expressed in bacterial cells; however, LKR was not. This was not due to lack of expression, because the LKR/SDH construct leads to the production of SDH but not LKR activity in bacteria. Moreover, the lack of production of active Arabidopsis LKR in bacteria was not due to a mutation in its sequence, because the same DNA produced active LKR when expressed in yeast cells. These results suggest that LKR may be activated by post-translational modification, which does not operate in prokaryotes. Indeed, we have recently found that the active LKR enzyme from soybean is a phosphoprotein and that removal of its phosphate residue(s) by alkaline phosphatase knocked out LKR activity in vitro (D. Miron, S. Ben-Ya'acov, H. Karchi, and G. Galili, submitted manuscript).

METHODS

Plant Material

Arabidopsis thaliana var C24 plants were grown in a greenhouse, and different tissues were collected from the developing plants for the isolation of the total RNA and in situ hybridization.

The cell culture of *Arabidopsis ecotype* Landsberg *erecta* was kindly provided by M.J. May (University of Oxford, Oxford, UK; May and Leaver, 1993). This culture was grown in MSMO liquid medium (Sigma), pH 5.7, containing 3% sucrose, 0.05 mg/L kinetin, and 0.5 mg/L naphthaleneacetic acid. The culture was placed on a rotary shaker at 110 rpm at 22°C in continuous fluorescent white light.

Cloning of the Full-Length cAt-LKR/SDH and cAt-SDH and Subcloning Them into Expression Vectors

The expressed sequence tag (EST) clone 23A3T7 and the λZAP II cDNA library (Kieber et al., 1993) were kindly provided by the Arabidopsis Biological Resource Center (Columbus, OH). To clone the full-length cAt-LKR/SDH from the λZAP II library, the cDNA from the EST clone was used as a probe to screen the library, as previously described (Sambrook et al., 1989). The plasmid containing the full-length cAt-LKR/SDH was excised from the λZAP II by using a helper phage, and its DNA sequence was determined by an automatic sequencer (model 373A, version 1.2.0; Applied Biosystems, Foster City, CA).

For expression of the putative monofunctional SDH in bacteria, an SmaI to XbaI DNA fragment containing the entire coding sequence of cAt-SDH was subcloned by a translational fusion into EcoRI (blunt ended with the Klenow fragment of DNA polymerase I) and XbaI sites of pUC18. For subcloning into the bacterial expression vector pET-15b, the coding sequence of cAt-SDH was excised with XbaI (blunt ended with the Klenow fragment) and SalI and subcloned as a translational fusion into the BamHI (blunt ended with the Klenow fragment) and XhoI sites of pET-15b to form the plasmid pET-15b-SDH.

For expression of the LKR/SDH sequence in bacteria, cAt-LKR/SDH was digested with EcoRI, which cleaves immediately after the LKR translation initiation codon (ATGAATTC). The plasmid was then blunt ended with the Klenow fragment, digested with NheI, which cleaves in the SDH domain, and subcloned into the NcoI (blunt ended with the Klenow fragment) and NheI sites of pET-15b-SDH, resulting in the plasmid pET-15b-cAt-LKR/SDH.

For expression in yeast, pET-LKR/SDH was digested with XbaI, which cleaves immediately upstream of the LKR translation initiation ATG codon, and PstI, which cleaves in the SDH domain. The insert was then inserted into the XbaI and PstI sites of pVT-102u, resulting in the plasmid pVT-102u-LKR.

Production of Recombinant Proteins in Bacteria and Yeast

The expression plasmids were transformed into *Escherichia coli* (Sambrook et al., 1989) and yeast cells (Ito et al., 1983) by using general heat shock and LiOAc transformation methods, respectively. Transformed bacterial cells were grown to mid-exponential phase (A_{600} of ~0.5 to 0.8) and then induced with 0.4 mM isopropyl β-D-thiogalactopyranoside for an additional 4 hr. Transformed yeast cells (mutant 8973b from A. Pierre, Université Libre de Bruxelles, Brussels, Belgium; Ramos et al., 1988) were grown to mid-log phase in liquid SC medium (Sherman et al., 1983) lacking uracyl.

Processing of Bacteria and Yeast for Analysis of LKR and SDH Activities

E. coli cells were precipitated, dissolved in one-tenth of buffer A (25 mM potassium phosphate buffer, pH 7.5, containing 1 mM EDTA, 1 mM DTT, and 10 μg/mL leupeptin), and sonicated. The total lysate was precipitated at top speed (16,000*g*) in a tabletop centrifuge for 10 min at 4°C, and the supernatant was used for activity assays. Yeast cells were precipitated, redissolved in one-tenth of buffer A, and broken by vortexing with glass beads for half an hour at 4°C. The lysate was precipitated again, and the supernatant was used for activity assays.

DNA Gel Blot Analysis

Extraction of genomic DNA was performed according to the procedure in Sambrook et al. (1989). DNA samples (10 μg) were electrophoresed in a 1% agarose gel and transferred to a Hybond N+ (Amersham) nylon membrane. The blots were hybridized for 12 to 16 hr at 65°C with ^{32}P-labeled probes containing either the LKR or SDH domain of cAt-LKR/SDH. Hybridization was performed in 5 × SSC (1 × SSC is 0.15 M NaCl, 0.015 M sodium citrate), 5 × Denhardt's solution (1× Denhardt's solution is 0.02% Ficoll, 0.02% PVP, and 0.02% BSA), and 1% SDS. Blots were washed twice for 10 min at 65°C in 1 × SSPE (1 × SSPE is 0.15 M NaCl, 10 mM sodium phosphate, 1 mM EDTA, pH 7.5) and 0.1% SDS, followed by another wash in 0.1 × SSPE and 0.1% SDS. Radioactive bands were detected by autoradiography. The hybridization probes included either the 1454-bp SalI-NdeI fragment of cAt-SDH (SDH probe) or a 771-bp NotI-HindIII fragment from cAt-LKR/SDH in pBluescript SK− (Stratagene, La Jolla, CA; LKR probe).

RNA Gel Blot Analysis

Total RNA was extracted from various tissues by using Tri-Reagent (MRC, Inc., Cincinnati, OH), according to the protocol provided by the manufacturer. RNA samples (20 μg) were electrophoresed in a 1% agarose gel containing 2.2 M formaldehyde and 50 mM 3-(*N*-morpholino) propanesulfonic acid, pH 7.0, and transferred to a Hybond N nylon membrane. Probe utilization, hybridization, and washing were as described above for the DNA gels blots. The migration of the 28S and 18S rRNAs was visualized by ethidium bromide staining of the gel before transfer to a membrane.

Partial Purification of the LKR and SDH from Arabidopsis Cell Culture

A 1-week-old cell culture was filtered, and the resulting cell pellet was frozen in liquid nitrogen and kept at −80°C until used. For purification, the frozen pellet was ground with a mortar and pestle and then homogenized using an Ultraturax (Ystral GmbH, Dottingen, Germany) in an equal volume of buffer A. After centrifugation at 25,000*g* for 15 min, the pH of the supernatant was brought to pH 5.6 with solid KH_2PO_4 and then fractionated with polyethylene glycol (PEG) 8000 between 7 and 14%. After fractionation with 14% PEG, the pellet was resuspended in one-tenth the initial volume of buffer A and loaded

onto an anion exchange DEAE-Sepharose column (Pharmacia). After washing the unbound protein, the column was eluted with a step gradient of 0 to 1 M KCl in buffer A.

Analysis of LKR and SDH Activities

The kinetics of LKR activity was measured spectrophotometrically by determining the rate of NADPH oxidation at 340 nm for 10 min at 30°C. The activity assays included 50 μg of protein extract in 0.3 mL of 0.1 M Tris-HCl, pH 7.4, 20 mM lysine, 14 mM α-ketoglutarate, and 0.4 mM NADPH. Each reaction also included a control lacking the substrate lysine. One unit of LKR was defined as the amount of enzyme that catalyzes the oxidation of 1 nmol of NADPH per min at 30°C.

The kinetics of SDH activity was measured spectrophotometrically by determining the rate of NAD^+ reduction at 340 nm for 10 min at 30°C. The activity assay included 50 μg of protein extract in 0.3 mL of 0.1 M Tris-HCl, pH 8.5, 2 mM saccharopine, and 2 mM NAD^+. Each reaction also included a control lacking the substrate saccharopine. One unit of SDH was defined as the amount of enzyme that catalyzes the reduction of 1 nmol of NAD^+ per min at 30°C.

Protein Determination

Protein levels were determined by the method of Bradford (1976), using the Bio-Rad protein assay kit.

In Situ Hybridization

For preparation of the hybridization probe, the LKR and SDH domains of cAt-LKR/SDH were subcloned separately into the pBluescript SK− plasmids. Digoxigenin-labeled sense and antisense probes were obtained by in vitro transcription using the digoxigenin RNA labeling kit (Boehringer Mannheim). Tissue preparation and in situ hybridization were conducted as described by Drews (1995). An antisense probe and the corresponding sense control probe were used in each experiment.

Primer Extension

Primer extension analysis was performed according to Sambrook et al. (1989), with several modifications. Total RNA (10 μg) from flowers was mixed with ^{32}P-end-labeled antisense primer located 20 to 46 nucleotides downstream of the transcription initiation ATG codon of cAt-SDH. The reaction was then incubated at 80°C for 10 min and cooled slowly to room temperature for annealing. Reverse transcription was conducted at 42°C for 1.5 hr. The reaction was stopped by boiling for 10 min and cooling on ice, and the mixture was then treated with RNase free of DNase for 30 min at 37°C. After ethanol precipitation, the primer extension product was analyzed on a sequencing gel along with a sequencing ladder of the same primer annealed to the relevant genomic fragment. Radioactive bands were detected by autoradiography.

Computer Analysis

DNA sequence analyses were performed using the Genetics Computer Group (Madison, WI) software package (version 8).

ACKNOWLEDGMENTS

We are indebted to Hua Zhang for excellent technical assistance. We also thank Dr. Hillel Fromm for critical reading of the manuscript and helpful comments; Dr. Shula Michaeli for help with the primer extension experiment; Dr. Joseph K. Ecker for making available the λZAP II cDNA library; the Arabidopsis Biological Resource Center for providing this library and the EST clone 23A377; Dr. Mike J. May for providing the Arabidopsis cell culture; and Dr. André Pierard for providing us with the yeast mutant 8973b. This work was supported in part by grants from the Israel Academy of Sciences and Humanities, National Council for Research and Development, Israel, and the Leo and Julia Forchheimer Center for Molecular Genetics. G.G. is an Incumbent of the Bronfman Chair of Plant Sciences.

Received May 5, 1997; accepted May 19, 1997.

REFERENCES

Arruda, P., and Da Silva, W.J. (1983). Lysine-ketoglutarate reductase activity in developing maize endosperm. Phytochemistry **22,** 2687–2689.

Bhattacharjee, J.K. (1985). α-Amino-adipate pathway for the biosynthesis of lysine in lower eukaryotes. Crit. Rev. Microbiol. **12,** 131–151.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. **72,** 248–254.

Brochetto-Braga, M., Lelte, A., and Arruda, P. (1992). Partial purification and characterization of lysine-ketoglutarate reductase in normal and opaque-2 maize endosperms. Plant Physiol. **98,** 1139–1147.

Bryan, J.K. (1980). Synthesis of the aspartate family and branched-chain amino acids. In The Biochemistry of Plants, Vol. 5, B.J. Miflin, ed (New York: Academic Press), pp. 403–452.

Dancis, J., Hutzler, J., Cox, R.P., and Woody, N.C. (1969). Familial hyperlysinemia with lysine ketoglutarate reductase insufficiency. J. Clin. Invest. **48,** 1447–1452.

Drews, G.N. (1995). Fixation and embedding for *in situ* hybridization with non-radioactive probes. In *Arabidopsis*, X.-W. Deng, R. Last, and D. Preuss, eds (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), pp. 1–32.

Falco, S.C., Guida, T., Locke, M., Mauvais, J., Sanders, C., Ward, R.T., and Webber, P. (1995). Transgenic canola and soybean seeds with increased lysine. Bio/Technology **13,** 577–582.

Feller, A., Dubois, E., Ramos, F., and Pierard, A. (1994). Repression of the genes for lysine biosynthesis in *Saccharomyces cerevisiae* is caused by limitation of lys14-dependent transcriptional activation. Mol. Cell. Biol. **14,** 6411–6418.

Fjellstedt, T.A., and Robinson, J.C. (1975). Purification and properties of L-lysine-α-ketoglutarate reductase from human placenta. Arch. Biochem. Biophys. **168,** 536–548.

Foster, A.R., Scislowski, P.W.D., Harris, C.I., and Fuller, M.F. (1993). Metabolic response of liver lysine α-ketoglutarate reductase activity in rats fed lysine limiting or lysine excessive diets. Nutr. Res. **13,** 1433–1443.

Galili, G. (1995). Regulation of lysine and threonine synthesis. Plant Cell **7**, 899–906.

Galili, G., Shaul, O., Perl, A., and Karchi, H. (1994). Synthesis and accumulation of the essential amino acids lysine and threonine in seeds. In Seed Development and Germination, H. Kigel and G. Galili, eds (New York: Marcel Dekker), pp. 811–831.

Ghislain, M., Frankard, V., Vandenbossche, D., Mattews, B.F., and Jacobs, M. (1994). Molecular analysis of the aspartate kinase-homoserine dehydrogenase gene from *Arabidopsis thaliana*. Plant Mol. Biol. **24**, 835–851.

Goncalves-Butruille, M., Szajner, P., Torigoi, E., Leite, A., and Arruda, P. (1996). Purification and characterization of the bifunctional enzyme lysine-ketoglutarate reductase-saccharopine dehydrogenase from maize. Plant Physiol. **110**, 765–771.

Hinnebusch, A.G. (1988). Mechanisms of gene regulation in the general control of amino acid biosynthesis in *Saccharomyces cerevisiae*. Microbiol. Rev. **52**, 248–273.

Ito, H., Fukuda, Y., Murata, K., and Kimura, A. (1983). Transformation of intact yeast cells treated with alkali cations. J. Bacteriol. **153**, 163–168.

Joshi, C.P. (1987). An inspection of domains between putative TATA box and translation start site in 79 plant genes. Nucleic Acids Res. **15**, 6643–6653.

Kalinowski, J., Cremer, J., Bachmann, B., Eggeling, L., Sahm, H., and Puhler, A. (1991). Genetic and biochemical analysis of the aspartokinase from *Corynebacterium glutamicum*. Mol. Microbiol. **5**, 1197–1204.

Karchi, H., Shaul, O., and Galili, G. (1994). Lysine synthesis and catabolism are coordinately regulated during tobacco seed development. Proc. Natl. Acad. Sci. USA **91**, 2577–2581.

Karchi, H., Miron, D., Ben-Yaacov, S., and Galili, G. (1995). The lysine-dependent stimulation of lysine catabolism in tobacco seeds requires calcium and protein phosphorylation. Plant Cell **7**, 1963–1970.

Kieber, J.J., Rothenberg, M., Roman, G., Feldman, K.A., and Ecker, J.R. (1993). *CTR1*, a negative regulator of the ethylene response pathway in Arabidopsis, encodes a member of the Raf family of protein kinases. Cell **72**, 427–441.

Markovitz, P.J., and Chuang, D.T. (1987). The bifunctional aminoadipic semialdehyde synthase in lysine degradation. J. Biol. Chem. **262**, 9353–9358.

Markovitz, P.J., Chuang, D.T., and Cox, R.P. (1984). Familial hyperlysinemias: Purification and characterization of the bifunctional aminoadipic semialdehyde synthase with lysine-ketoglutarate reductase and saccharopine dehydrogenase activities. J. Biol. Chem. **259**, 11643–11646.

May, M.J., and Leaver, C.J. (1993). Oxidative stimulation of glutathione synthesis in *Arabidopsis thaliana* suspension cultures. Plant Physiol. **103**, 621–627.

Moller, B.L. (1976). Lysine catabolism in barley (*Hordeum vulgare* L.). Plant Physiol. **57**, 687–692.

Noda, C., and Ichihara, A. (1978). Purification and properties of L-lysine-α-ketoglutarate reductase from rat liver mitochondria. Biochim. Biophys. Acta **525**, 307–313.

Ogawa, H., and Fujioka, M. (1978). Purification and characterization of saccharopine dehydrogenase from baker's yeast. J. Biol. Chem. **253**, 3666–3670.

Ramos, F., Dubois, E., and Pierard, A. (1988). Control of enzyme synthesis in the lysine biosynthetic pathway of *Saccharomyces cerevisiae*. Eur. J. Biochem. **171**, 171–176.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, 2nd ed. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Sherman, F., Fink, G., and Hicks, J.B. (1983). Methods in Yeast Genetics: A Laboratory Manual. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Shotwell, M.A., and Larkins, B.A. (1988). The biochemistry and molecular biology of seed storage proteins. In The Biochemistry of Plants, Vol. 15, B.J. Miflin, ed (New York: Academic Press), pp. 297–345.

Zhu-Shimoni, J.X., Lev-Yadun, S., Matthews, B., and Galili, G. (1997). Expression of an aspartate kinase homoserine dehydrogenase gene is subject to specific spatial and temporal regulation in vegetative tissues, flowers, and developing seeds. Plant Physiol. **113**, 695–706.

Evidence Appendix E

Doerks (TIG14, No. 6:248-250, June 1998)

This was cited by the Examiner in the Office Action mailed on January 25, 2007.

Protein annotation: detective work for function prediction

Computer analysis of genome sequences is currently one of the essential steps for obtaining functional and structural information about the respective gene products. Database searches are used to transfer functional features from annotated proteins to the query sequences. With the increasing amount of data, more and more software robots perform this task[1]. While robots are the only solution to cope with the flood of data, they are also dangerous because they can currently introduce and propagate misannotations[2,3]. On the one hand, functional information is often only partially transferred (underprediction). For example, information is not usually extracted for each functional unit (protein domain) but just taken from the one-line description of the best database match (so multifunctionality is rarely considered). On the other hand, overpredictions are common because the highest-scoring database protein does not necessarily share the same or even similar functions.

Definition and collection of uncharacterized protein families

To avoid unnecessary propagation of poor annotation, we have collected putative, poorly annotated proteins that are usually labeled as 'hypothetical' or just as 'ORF' (open reading frame). We operationally defined uncharacterized protein families (UPFs) to be families of proteins that: (1) contain members in at least three taxonomically distinct (and phylogenetically 'distant') species; and (2) do not contain (to the best of our knowledge) biochemically characterized proteins.

A collection and classification of these proteins should allow: (a) utilization of family information and thus a more detailed characterization; (b) simplification of update procedures for the entire families if functional information becomes available for at least one member; and (c) a careful annotation of functional features that avoids the pitfalls described above.

As the numerous genome sequencing projects progress, more and more of these UPFs emerge in sequence databases. We gave high priority to families that contain members in at least two of the three major kingdoms (archae, eubacteria, eukaryotes). The original 'family' definition was based on significant hits in the statistics provided by FASTA (Ref. 4) or gapped BLAST (Ref. 5).

Annotation of UPFs in SWISS-PROT and PROSITE databases

A serial number has been assigned to each UPF and to each of the corresponding SWISS-PROT (Ref. 6) entries. A SWISS-PROT document file lists all the current UPFs and their members in SWISS-PROT. This document is available on the WWW (Ref. 7). In the majority of cases, PROSITE entries[8] have already been created to document the respective family. Whenever a member of a UPF family is biochemically characterized, that family ceases to be considered as a UPF and is deleted from the list. However, information is provided that allows its history to be traced. For example:

Family: UPF0002 IDELETED1
Taxonomic range: Eubacteria
Comments: Now characterized as a family of pseudouridylate synthases (EC 4.2.1.70).
Prototype: RSUA_ECOLI (Accession No. P33918)
PROSITE entry: PDOC00685

Function prediction for the UPFs

The annotation is handled rather conservatively (see below) because functional overpredictions are most dangerous given the many opportunities for error propagation in sequence databases[2,3]. Nevertheless, we intended to retrieve as many functional features as possible for each UPF using comparative analysis. Thus, each UPF was subjected to a variety of sequence analysis methods[9]. In brief, several members of each UPF were compared with a database of non-identical protein sequences, daily updated at the EMBL using PSI-BLAST (Ref. 5) with a conservative expected ratio of false positives (E = 0.001) as a threshold for each iteration. Sequences were preprocessed by filtering for transmembrane[10] and coiled-coil regions[11]. A multiple alignment was constructed for each UPF using ClustalX (Ref. 12). If PSI-BLAST did not identify a relationship to characterized proteins, other iterative methods such as Wisetools (Ref. 13) and Most (Ref. 14) were applied. They also use family information, that is, give more weight to conserved positions and so on, but have the advantage that the underlying multiple alignments can be checked and improved manually (on the cost of speed and the 'easy to use' feature).

Finally, all searches were repeated using a sequence database that only contained sequences from entirely sequenced genomes to reduce noise effects[2,15]. For example, PSI-BLAST E-values depend on the database and a database match might be significant using a small database but becomes insignificant if more background noise (unrelated or redundant sequences) is added.

In many cases, the iterations revealed the relationship of the UPFs with other proteins, families or superfamilies. As the main focus here was to assign functional features, the iterations have not been continued when a reasonable prediction could be made. Criteria for the latter were matches to known active site patterns or conserved motifs resembling those in PROSITE as well as the positioning of UPF members within phylogenetic trees. Transmembrane regions were identified in 13 (22%) of the 58 UPFs, although functional predictions for these 13 have not been made. Of the remaining 45 UPFs, 25 could be related to proteins with annotated functional features (Table 1).

Pitfalls in function assignments

The predictions required careful inspection of the functional annotations of the matched database proteins. To illustrate the difficulties, Table 2 shows the result of a Blast search for UPF0002 that includes quite a few proteins with annotations (in addition to the first hits that are labeled as 'hypothetical'). Only one can give a clue about functional features; others are simply wrong, misleading or uninformative.

Another typical assignment error is caused by the sequence similarity of the query to a region that is independent from the one that was the basis for the annotation. For example, the hypothetical protein HI0722 (Accession No. P44842, ID: YJGZ_HAEIN), a member of the UPF0020 family, shows significant similarity to two proteins (GenBank entries gi|2314679 and gi|2698141) in *Helicobacter pylori* and *Borrelia burgdorferi*, respectively, which are wrongly annotated as proline dipeptidases (pepQ). The annotation is based on the N-terminal homology of these two proteins with the C-terminal region of proline dipeptidase (pepQ) (gi|42358) of *E. coli*, which does not harbor the catalytic activity of this enzyme.

There were even examples in which homologs scored best in PSI-BLAST (Ref. 5) that did not have the same catalytic activity because active site residues of the characterized family were not conserved. However, there were significantly lower scoring homologs with perfect matches of their (distinct) catalytic site residues to the query. For example, the UPF0046 family has clear amino acid similarity to proteases that are easily found by PSI-BLAST (Ref. 5) in the fourth iteration; yet, residues involved in metal-binding are only shared with a purple acid phosphatase family that is only picked up in the ninth iteration. The E-value of 1e-5 compared with proteases (E-value of 5e–78) remain considerably higher in subsequence iterations. Such instances have

PII: S0168-9525(98)01486-5

implications for current function prediction programs in which the function of the best hit is transferred. Clearly, another generation of methods is required that include checks for the presence of functionally important residues.

Use of phylogenetic trees

As most of the database proteins with functional annotations were only distantly related to members of the UPFs, transfer of functional information is extremely difficult and arbitrary. The majority of UPFs turned out to be related to enzymes, and based on the conservation of the active site residues one can assume that at least the basic catalytic mechanism remains the same. This, however, is of little predictive value as some families, e.g. those with the α/β hydrolase fold collected in SCOP (Ref. 16) are huge and harbor numerous distinct catalytic activities, such as lipases, esterases, dehalogenases, peptidases, peroxidases and lyases. We have therefore constructed phylogenetic trees of selected members of the UPFs and of related, but distinct families that have been identified during the analysis (Fig. 1). On some occasions, the UPF members clearly clustered with proteins that all performed the same function (Fig. 1a), but in most of the cases the UPFs were of equal distance to distinct enzymatic activities (Fig. 1b), thus not allowing any detailed predictions.

Although the studied protein families were bound to be difficult for function predictions because a considerable number of teams were unable to find functional

TABLE 1. Predicted functional features for 25 UPFs

UPF No.	Family size[a]	Predicted function
02	70	Pseudouridylate synthase
04	60	Methyltransferase
07	15	Cytidyltransferase[b]
08	30	ATPase
09	40	GTPase
10	10	Aldose 1-epimerase
11	10	Methyltransferase[b]
12	25	Nitrilase
17	30	Hydrolase
19	15	Phosphate-binding protein (TIM BARREL)
20	40	N6-adenine-specific methylase
21	50	ATPase
26	30	Two domain protein : iron/sulfur binding and amidotransferase
30	10	Amidotransferase
31	30	Sugar kinase
34	20	Pyrimidin-binding oxidoreductase (TIM BARREL)
35	20	Mutator mutt protein (7,8-dihydro-8-oxoguaninetriphosphatase)
36	70	Hydrolase
37	10	Oxydoreductase
38	35	ATPase[b]
42	10	ATPase
46	15	Phosphatase
49	50	N6-adenine-specific methylase
53	40	CBS domain protein
55	10	Glutaredoxin

[a]The numbers of family members are approximate because of daily changes in databases and loose family definitions.
[b]*E. coli* member also predicted by Koonin *et al.*[17] (UPF0007: nucleotidyltransferase).
Abbreviation: UPFs, uncharacterized protein families.

TABLE 2. Misleading annotations: PSI-BLAST results for the UPF0002 family (first iteration)

Ranking	Annotation	Probability	Commentary
1	Gnl|PID|e332795 (Z98268) hypothetical protein MTCI25.35 [Mycobacterium tuberculosis]...	(2e–75)	
4	**Sp|P33643|SFHB_ECOLI SFHB PROTEIN**	(1e–67)	SFHB is a gene name (suppressor of the temperature-sensitivity of *ftsb1* mutation) and does not give much functional insight
5	Gnl|PID|e1185138 (Z99112) alternative gene name: *ylmE*; similar to hypothetical proteins [Bacillus subtilis]...	(3e–65)	
37	**Sp|Q12362|RIB2_YEAST DRAP DEAMINASE** >gi|1078332|pir||S50972 RIB2 protein – yeast (Saccharomyces cerevisiae) >gi|642221 (Z21618) DRAP deaminase [Saccharomyces cerevisiae] >gi|1419887|gnl|PID|e252279 (Z74808) ORF YOL066c [Saccharomyces cerevisiae]...	(7e–50)	The homology is not in the catalytic region and does not hold for other deaminases
40	**Sp|P33918|RSUA_ECOLI 16S PSEUDOURIDYLATE 516 SYNTHASE (16S PSEUDOURIDINE 516 SYNTHASE) (URACIL HYDROLYASE)**	(2e–48)	Function prediction based on this protein
41	**Sp|Q47417|YQCB_ERWCA EXOENZYME REGULATION REGULON ORF1** >gi|1626643|pir||S45107 hypothetical protein 1 – Erwinia carotovora >gi|496598 (X79474) ORF1 [Erwinia carotovora]...	(7e–48)	Misleading annotation, operon architecture is not conserved between species

Annotations that hamper functional predictions illustrated by the example of the UPF0002 family. Based on the recent experimental characterization of pseudouridylate synthase[18], this family has been deleted from the UPF list (see text). Nevertheless, the various, partly contradictory annotations (bold) are extremely difficult to parse for automatic function prediction programs. For brevity, the PSI-BLAST results have been cut (...).

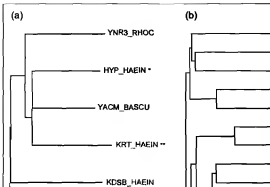


FIGURE 1. (a) Phylogenetic trees of selected members of UPF0007 that indicate a likely function as UPF0007 members with cytidyltransferase activities (red) and related uridilyltransferases (blue) are more divergent (*pir database entry, pir|s64156; **pir database entry, pir|s49238). (b) No clear enzymatic activity can be predicted for UPF0017 members: They clearly have the hydrolase fold but have equal distance to peroxidases (red), esterases (green), peptidases (blue) and other hydrolases (pink) (***GenBank entry gi|1001804). The trees were calculated using CLUSTALX (Ref. 12).

features therein, it is noteworthy that there was not a single case in which we were able to predict the precise mechanism and the substrate specificity. Nevertheless, the information about an enzymatic activity and the likely reaction mechanism of the 25 UPFs should prove useful for the analysis of upcoming genome sequences.

Annotation with the right level of precision helps in future projects

In summary, we were able to provide some functional annotation for more than 700 of about 1300 proteins clustered in 25 of the 58 distinct UPFs. Most of them are currently named 'hypothetical protein' so that their annotation adds enormous value to these sequences. For another 13 UPFs currently containing about 250 proteins, the presence of transmembrane regions was recorded. This annotation is now being incorporated into PROSITE and SWISS-PROT so that these features can be assigned to newly sequenced genes as well.

The difficulties we faced in assigning functions by sequence similarity also indicate that many of the automatic predictions by most of the software robots are probably erroneous. Because of the current policies of most of the sequence databases, correction of annotations is very hard to realize. Thus, there should be a combined effort by the database teams, the authors of the current entries, and the community, to work towards a careful functional annotation of all the sequences that become publicly available.

References

1 Boguski, M. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71
2 Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 425–427
3 Bhatia, U., Robinson, K. and Gilbert, W. (1997) *Science* 276, 1724–1725
4 Pearson, W.R. and Miller, W. (1992) *Methods Enzymol.* 210, 575–601
5 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
6 Bairoch, A. and Apweiler, R. (1998) *Nucleic Acids Res.* 26, 38–42
7 httpd://www.expasy.ch/cgi-bin/lists?upflist.txt
8 Bairoch, A., Bucher, P. and Hofmann, K. (1998) *Nucleic Acids Res.* 25, 217–221
9 Bork, P. and Gibson, T. (1996) *Methods Enzymol.* 266, 162–184
10 Von Heijne, G. (1992) *J. Mol. Biol.* 225, 487–494
11 Lupas, A., van Dyke, M. and Stock, J. (1991) *Science* 252, 1162–1164
12 Thompson, J.D. *et al.* (1997) *Nucleic Acids Res.* 25, 4876–4882
13 Birney, E., Thompson, J.D. and Gibson, T.J. (1996) *Nucleic Acids Res.* 24, 2730–2739
14 Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 91, 12091–12095
15 Bork, P. and Koonin, E.V. (1998) *Nat. Genet.* 18, 313–318
16 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *Mol. Biol.* 247, 536–540
17 Koonin, E., Mushegian, A., Galperin, M. and Walker, D. (1997) *Mol. Microbiol.* 25, 619–637
18 Wrzesinski, J. *et al.* (1995) *Biochemistry* 34, 8904–8913

Tobias Doerks
doerks@embl-heidelberg.de

EMBL, Meyerhofstrasse 1, 69012 Heidelberg and Max-Delbruck-Center for Molecular Medicine, Berlin-Buch, Germany.

Amos Bairoch
amos.bairoch@medecine.unige.ch

Swiss Bioinformatics Institute and University of Geneva, Switzerland.

Peer Bork
bork@embl-heidelberg.de

Evidence Appendix F

Smith et al. (Nature Biotechnology 15:P1222-1223, November 1997)

This was cited by the Examiner in the Office Action mailed on January 25, 2007.

BIOINFORMATICS

The challenges of genome sequence annotation or "The devil is in the details"

Temple F. Smith and Xiaolin Zhang

Two powerful, competing pressures are acting on various genome sequencing projects: One, to release new sequences as quickly as possible; and two, to provide them with maximally complete and accurate annotation. This rather incongruent combination has led to a strong interest in developing efficient and accurate automated, large-scale sequence annotation procedures.

There have, in fact, been a number of attempts in both industry and academia to speed new sequence annotation. In their simplest form, these have been little more than post-processors acting on standard high-speed sequence similarity search tools such as BLAST. These post-processing assigns the annotation from the best-matched previously known sequence to each new sequence.

This is, of course, a generalization of successful approaches used by many researchers to assign probable functions to new sequences when previously studied and recognizable homologs exist. However, when applied in an automated manner to large data sets with minimum review, such approaches can lead to serious degradation of the wealth of incoming genomic data.

There are more problems with the simple best match functional annotation inheritance (BMAI) than the two traditionally recognized, those being the assessing of biological significance in terms of match statistical significance, and the choice between the sensitivity of the very fast, but approximate, sequence similarity search algorithms and the mathematically rigorous, but much slower, optimal algorithms.

In the first place, it is easy to assign various measures of confidence to new annotation based on match statistics, and there is good evidence that approximate maximum similarity tools such as BLAST do nearly as well as any of the slower, full dynamic programming methods. Second, the newer versions of BLAST have high sensitivity, identifying local sequence pairwise similarities, including alignment gaps. The inclusion of alignment gaps was one of the main advantages of the slower dynamic programming methods.

Temple F. Smith is director and Xiaolin Zhang is a research associate at the BioMolecular Engineering Research Center, Boston University College of Engineering, 36 Cummington Street, Boston, MA 02215 (tsmith@darwin.bu.edu).

No, the major problems associated with nearly all of the current automated annotation approaches are—paradoxically—minor database annotation inconsistencies (and a few outright errors). This is particularly true for the large and often complex protein families. Why are these the major problems, rather than the two more obvious ones previously mentioned?

Clearly, for researchers studying a particular protein family, most database annotation inconsistencies make little difference in the search for new, even distant members. A local expert either knows the range and/or history of the annotation terminology used by colleagues in different subfields, or perhaps more importantly, the expert will spend the time to backtrack apparent inconsistencies.

Even in those cases involving structurally complex proteins composed of multiple domains, all of which may not be fully or properly annotated, the expert generally carefully dissects matches to distinct domains, and backtracks each domain's annotations. However, in the large-scale genomic projects, having a local expert to work on each protein family is not an option. Yet the integration of genomic information across multiple protein families, multiple fields of expertise and taxa, is just what is envisioned to form the foundations of the next century's biology and biotechnology.

The basic problem of inconsistent nomenclature arises largely because sequence information and its annotation derives from many diverse subdivisions of the biological sciences during a time of rapid change in our understanding. In an emerging field such as molecular biology, let alone "comparative genomics," strictly controlled vocabularies would not only be difficult to impose, but are probably undesirable! The evolution and refinement of the vocabulary is an anticipated outcome of our increasing knowledge.

Some inconsistencies are simple, such as the reference to tRNA synthetases in fungi as tRNA ligases (which of course they are) or the use by Americans and most Europeans of dihydroxyacetone-P for a glycolytic intermediate that the Japanese and English generally call glycerone-P. There are many cases of equivalent, but different, terminology. For example, in the well-studied G protein case, among 27 distinct G β-subunit GenBank/SWISS-PROT entries, there are 18 different protein names or keyword sets. A list of synonyms can be constructed in such cases, some of which will be species or field specific.

There are numerous cases in which proteins of very different current functions are homologous in that they evolved from a common ancestor and will match with significant sequence similarity. For example, numerous proteins sharing multiple WD-repeats have been labeled transducin-like or transducin homologs, yet share no common signal transduction function[1]. The rather widespread improper use of synthetase for synthase and the converse, however, cannot be fixed by a thesaurus, since whether the enzyme in question requires ATP or not is not a matter of alternate terminology. Without the careful use of synonym tables in

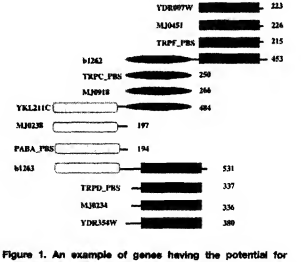


Figure 1. An example of genes having the potential for annotation inheritance transitivity. The three two-domain proteins, b1262, YKL211C, and b1263, share no single domain in common. Domains are labeled by colors: red, N-phosphoribosyl anthranilate isomerase; green, indole-3-glycerol phosphate sythhase; yellow, anthranilate synthase, subunit II; blue, anthranilate phosphoribosyltransferase.

combination with review of commonly missed terminology, any simple BMAI approach will often end up propagating the less desirable or erroneous annotations.

Random propagation of faulty annotation, however, is only the tip of the annotation problem iceberg. In the case of multidomain proteins, most simple BMAI approaches will at best annotate only the most similar of the domains, and at worst will attach the annotation of a nonshared domain from the matched protein.

The first of these, incomplete annotation, is seen in the recently released *Escherichia coli* genome data for ORF b1262, a 453-residue, multifunctional protein[2]. Here, the first 253 amino acid residues comprise the indole-3-glycerol phosphate synthase domain. This matches single-domain homologs in *Methanococcus jannaschii* and *Bacillus subtilis* and the carboxy-terminal domain of the protein product of one yeast gene, YKL211C. The second domain of the *E. coli* protein residues 259 through 443 matches the *N*-phosphoribosyl anthranilate isomerase, single-domain protein in *M. jannaschii*, *B. subtilis*, and yeast (and this function is currently unannotated).

An incorrect inheritance via a matched multidomain protein is seen in the *M. jannaschii* ORF pair, MJ0234 and MJ0238. Both match the *E. coli* ORF b1263, a bifunctional enzyme of two separate domains. Both *M. jannaschii* genes have been annotated, however, by only one of the two functions: anthranilate synthase subunit II, which is

What must be done to avoid continued annotation inconsistency, incompleteness, and erroneous propagation?

associated only with the first 176 of b1263's 531 amino acids, and that region is matched only by MJ0238 (Fig. 1).

What must be done to avoid continued annotation inconsistency, incompleteness, and erroneous propagation? First, any automation must be rather sophisticated. It must, for a start, recognize large differences in the length of matching sequences; it must associate annotation with specific subsequences; it must recognize all differences among the annotations of the homologs to the matched sequence; and, whenever possible, sequence similarity should be identified via shared conserved sequence patterns or profiles that have been carefully annotated, consistent with the entire family characterized by that pattern. All approaches should exploit the best available synonym tables, such as those available through resources like PROSITE, the Enzyme Commission, or the US National Library of Medicine's UMLS database. Finally, any annotation strategy must be designed to support an evolving nomenclature and rapidly expanding knowledge base.

Even if it takes an extended period of time to annotate the new genome data more carefully and completely now, it will surely be more cost effective than redoing it later. Recall that the correcting and/or updating of all of the historical data in largely archival sequence databases such as GenBank or SWISS-PROT, has not yet been completed—probably for good reasons of cost and time. We in the basic research and biotechnology communities must not let our excitement or our impatience for the new data degrade its annotation and longer-term utility.

1. Neer, E.J., Schmidt, C.J., Nambudripad, R., and Smith, T.F. 1994. The ancient regulatory-protein family of WD-repeat proteins. *Nature* 371:297–300.

2. ORF data obtained from:
M. jannaschii: www.tigr.org/tdb/mdb/mdb.html
E. coli: www.genetics.wisc.edu
S. cerevisiae: speedy.mips.biochem.mpg.de
B. subtilis: www.pasteur.fr/bubdi/subti_fiat.html

Evidence Appendix G

Brenner (TIG 15, 4:132-133, April 1999)

This was cited by the Examiner in the Office Action mailed on January 25, 2007.

Errors in genome annotation

At the time that Watson and Crick proposed a structure for DNA, a visionary might have suggested that the complete genetic sequence of an organism would eventually be known. However, nobody could have realistically proposed that machines could automatically indicate gene functions. Yet precisely this has been achieved: with no laboratory experiments at all, the roles of most genes in several organisms have been reported.

But how reliable are these functional assignments, upon which we depend for understanding genes and genomes? Without laboratory experiments to verify the computational methods and their expert analysis, it is impossible to know for certain. However, a simple procedure can place a rough upper bound on their accuracy. I have compared three different groups' functional annotation[1–3] for the *Mycoplasma genitalium* genome[4] (Fig. 1). Where two groups' descriptions are completely incompatible, at least one must be in error. In my analysis, there is no penalty for vague or absent functional assignment. Furthermore, I always assume that as many groups as possible have the right description (Fig. 2).

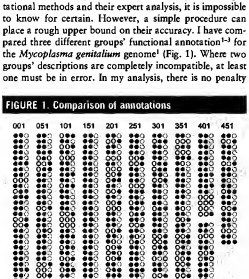
The results are disappointing for those expecting reliable annotation (Table 1). *M. genitalium* was reported to have just 468 genes, many of which are fundamental for all life and therefore easy to analyse. Nonetheless, the error rate is at least 8% for the 340 genes annotated by two or three groups. This value may not be uniform across the three groups, nor does it reflect the overall significance of a group's results. Genes annotated by only one group were not considered, but include such improbable bacterial functions as B-cell enhancing factor, mitochondrial polymerase, and serotonin receptor. This analysis cannot detect those cases where multiple groups arrived at consistent but wrong conclusions – a likely occurrence because all relied on similar methods and data. This evaluation also ignores minor disagreements in annotation, and disparities in degree of specificity (possibly indicating problematic overprediction of function[5]). Therefore, the true error rate must be greater than these figures indicate.

There are several possible reasons why the functional analyses have mistakes, as described at greater length elsewhere[6–8]. For example, it may be that the similarity between the genomic query and database sequence is insufficient to reliably detect homology, an issue solvable by appropriate use of modern and accurate sequence comparison procedures[9,10]. A more difficult problem is accurate inference of function from homology. Typical database searching methods are valuable for finding evolutionarily related proteins, but if there are only about 1000 major superfamilies in nature[11,12], then most homologs must have different molecular and cellular functions.

The annotation problem escalates dramatically beyond the single genome, for genes with incorrect functions are entered into public databases[8]. Subsequent searches against these databases then cause errors to propagate to future functional assignments. The procedure need cycle only a few times without corrections before the resources that made computational function determination possible – the annotation databases – are so polluted as to be almost useless. To prevent errors from spreading out of control, database curation by the scientific community will be essential[8,13].

To ensure that databases are kept usable, the intent of a gene annotation should be clear: does it indicate homolog, ortholog, and/or functional equivalence? Fortunately, some databases already incorporate this information explicitly (e.g. Ref. 14). Errors will, of course, still creep in. To help eliminate the collateral damage, computational assignments should clearly be flagged as such, and they should also indicate their source (which would allow propagation of corrections) and a measure of confidence in their accuracy. This will require new research and development in algorithms and databases, and a broad commitment to maintaining these resources. In short, the accessible documentation needed for reproducibility of a computational function determination should be commensurate with that for a corresponding laboratory bench experiment.

FIGURE 1. Comparison of annotations



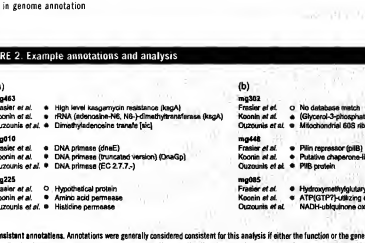
Three dots represent (left to right) Fraser *et al.*[2], Koonin *et al.*[3] and Ouzounis *et al.*[1] annotations for each of the 468 *M. genitalium* genes. (Tentative cases from Ouzounis *et al.*[1] were not used.) An open black circle indicates lack of a substantial functional annotation. Compatible annotations are colored identically, while conflicting annotations are in different colors. It is unknown which, if any, of the annotations are actually correct. There are 300 cases where Ouzounis *et al.*[1] simply reported the SWISS-PROT annotation of the same *M. genitalium* gene, indicated by colored open circles. Because Fraser *et al.*[2] annotation played a role in SWISS-PROT descriptions, these Ouzounis *et al.*[1] annotations were not included in this analysis. Though not incorporated in Table 1, the color indicates the compatibility of the functional annotation. The conflict/compatibility analysis here is itself certain to have errors; however, these should not affect the magnitude of the measured annotation error rate.

Steven E. Brenner
brenner@hhpcr.stanford.edu

Department of Structural Biology, Stanford University, Fairchild Building, Stanford, CA 94305-5126, USA.

0168-9525/99/$ – see front matter PII: S0168-9525(99)01706-0

FIGURE 2. Example annotations and analysis



(a) **Consistent annotations.** Annotations were generally considered consistent for this analysis if either the function or the gene name match (e.g. mg463; mg010). An exception is when one group uses a gene name and another specifically notes that the current gene is a paralog and not identical (consider mg010). Where the descriptions from different groups were compatible, but of different levels of specificity, this was considered a correct assignment (e.g. mg225). The difficulty of reconciling pairs of descriptions to determine whether they reflect compatible functions makes this analysis imprecise. Generally, the approach here is generous and should err on the side of detecting too few errors; it is usually more permissive than Ref. 5. **mg463:** Frasier *et al.*[1] and Koonin *et al.*[2] describe different aspects of function, but give the same gene name. The Ouzounis *et al.*[3] description is compatible with that from Koonin *et al.*[2], but less specific. All three annotations are considered correct for this analysis. **mg010:** Frasier *et al.*[1] and Ouzounis *et al.*[3] agree that this is a DNA primase. Koonin *et al.*[2] use a different gene name and explicitly state that this is a truncated protein. Because of the common functional descriptions, all three are considered correct. However, if Koonin *et al.*[2] had been more explicit in indicating a functional difference, then their annotation would have been marked as conflicting. (Note that mg256) is also annotated as a DNA primase by all three groups.) **mg225:** the Ouzounis *et al.*[3] annotation of histidine permease is more specific than the Koonin *et al.*[2] description of amino acid permease. It may be that histidine permease is an (incorrect) overprediction of function, or it could be correct. The two annotations are considered consistent, and the decision of Frasier *et al.*[1] not to provide a function is not penalized. (b) **Inconsistent annotations. mg302:** lack of a functional assignment from Frasier *et al.*[1] is not penalized. The Koonin *et al.*[2] and Ouzounis *et al.*[3] annotations are wholly inconsistent. This leads to a conflict and a minimum error rate of 50%. Note that the assessment methodology also behaves correctly when two annotators provide different functions for a multi-functional enzyme: each of the annotators is half right and half wrong, and the assessment assigns a 50% error rate. **mg448:** Frasier *et al.*[1] and Ouzounis *et al.*[3] both describe the gene as *pilB*. The encoded protein is involved in pilin formation, and its biochemical function is catalysis of methionine sulfmide oxidation/reduction in proteins. The Koonin *et al.*[2] annotation, chaperone-like protein, could conceivably be compatible but this is not likely. Because of uncertainty regarding compatibility of the Koonin *et al.*[2] annotation and its qualification as putative, this set of annotations is right on the threshold of consideration. For this analysis, the Koonin *et al.*[2] annotation was considered to be in conflict with the others, giving a minimum error rate of 33%. **mg085:** all three groups provide contradictory functions. The function described by Frasier *et al.*[1] of HMG-CoA reductase is EC 1.1.1.34, while the NADH-ubiquinone oxidoreductase annotated by Ouzounis *et al.*[3] (nuo6n_marpa) is EC 1.6.5.3. Neither enzyme uses ATP or GTP, as specified by Koonin *et al.*[2] The analysis assumes one is correct and marks two incorrect. Note: Ouzounis *et al.*[3] annotations equivalent to SWISS-PROT included in these examples are not included in the Table 1 analysis.

Acknowledgements

A previous version of this analysis was performed at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK. M. Levitt, C. Chothia, B. Al-Lazikani and P. Koehl provided stimulating discussion.

TABLE I. *M. genitalium* annotations, conflicts and error rates

No. groups annotating gene	No. genes	Annotations per group* Frasier *et al.*[1]	Koonin *et al.*[2]	Ouzounis *et al.*[3]	Total annotations	No. conflicts	Minimum error rate
0	33	–	–	–	–	N/A	N/A
1†	95	14	15	66	95	N/A	N/A
2	318	279	317	40	636	45	7%
3	22	22	22	22	66	10	15%
Sum (2+3)	340	301	339	62	702	55	8%

Summary of annotations made by each group (Fig. 1), minimum number of conflicting annotations (see Fig. 2), and the resulting minimal fraction of annotations that are erroneous.
*Frasier *et al.*[1] data from http://www.tigr.org/tdb/mdb/mgdb/mgdb.html. Koonin *et al.*[2] data from http://www.ncbi.nlm.nih.gov/Complete_Genomes/Mgen. Ouzounis *et al.*[3] data from http://www.embl-heidelberg.de/~genequiz/mycogen.new.html. Instances where Ouzounis *et al.*[3] reported SWISS-PROT annotation of the same gene were removed to avoid duplication with Frasier *et al.*[1] entries. However, even if all of these 300 annotations are included, the minimum annotation error rate drops only to 6%. All annotations were collected in 1996, shortly after the genome was released.
†No comparative analysis is possible when only one group made an annotation.

References

1 Frasier, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403
2 Koonin, E.V. *et al.* (1996) Sequencing and analysis of bacterial genomes. *Curr. Biol.* 6, 404–416
3 Ouzounis, C. *et al.* (1996) Novelties from the complete genome of *Mycoplasma genitalium*. *Mol. Microbiol.* 20, 898–900
4 Overbeek, T. *et al.* (1993) Protein annotation: detective work for function prediction. *Trends Genet.* 14, 248–250
5 Galperin, M.Y. and Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.* 1, 7
6 Smith, T.F. and Zhang, X. (1997) The challenges of genome sequence annotation or 'The devil is in the details'. *Nat. Biotechnol.* 15, 1222–1223
7 Bork, P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707–725
8 Bork, P. and Bairoch, A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.* 12, 425–427
9 Brenner, S.E. *et al.* (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078
10 Altschul, S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.* 6, 119–129
11 Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544
12 Brenner, S.E. *et al.* (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369–376
13 Smith, T.F. (1998) Functional genomics – bioinformatics is ready for the challenge. *Trends Genet.* 14, 251–329
14 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637

Evidence Appendix H

Borks (TIG12, 10:425-427, October 1996)

This was cited by the Examiner in the Office Action mailed on January 25, 2007.

Genetwork

Go hunting in sequence databases but watch out for the traps

The large amount of data created by world-wide sequencing efforts calls for automation in data handling and analysis. This requires accurate storage and updating mechanisms as well as appropriate retrieval software. User-friendly interfaces are also needed, as the number of researchers that access the information stored in public sequence databases is increasing considerably. Although the database teams are aware of the demands and the invaluable sequence databases are improving, they are also the product of history and, like the accessing software, far from perfect. Thus, at present, working with sequence databases requires knowledge about their powers and their pitfalls. Here, we concentrate on some of the problems that many users are unaware of, but that can have a considerable influence on the interpretation of the data. Some of the more frequent problems are summarized below, and some specific examples are given in Boxes 1 and 2.

Problems within the sequence databases themselves

Sequencing errors seem to be in the order of 0.1% (Ref. 1) (excluding ESTs, single reads with a very high error rate) affecting about 5% of the proteins[2]. When screening 500 human proteins in SWISS-PROT that have been published separately more than once, we find that 0.3% of the amino acids are different, this is a lower limit as lots of corrections have already been done and as sequences appearing in two different publications are often not independent. In any case, only frameshift errors and artificial stop codons can be detected unambiguously; point mutations are hard to verify as natural polymorphism or strain differences cannot easily be excluded. Even if the rate seems low, errors can accumulate in the sequence of interest (see Ref. 3 for an example) and can lead to functional misinterpretations. Moreover, although the quality of sequencing is improving, budget calculations might favor quantity instead of quality in the near future; the successful strategies based on ESTs demonstrate that data quality and its interpretation remain a major issue. Errors of various sources are also a major problem for other molecular databases such as Brookhaven Protein Data Bank[4].

The processing of raw DNA, by database management software is another serious source of problems. For example, false translation of genomic DNA into gene products, having missed exons or translated introns, leads to erroneous entries in protein sequence databases; the correct initiating methionine is not always chosen as a translation start; or ORFs translated from the opposite strand of the gene end up as proteins. The challenge is to improve prediction methods as the widely used algorithms of gene identification in higher eukaryotes have only an accuracy between 60–70% (Ref. 5): nearly a third of the automatically predicted proteins from genomic DNA without clear homologs are expected to contain some serious errors.

Erroneous annotation is also common, ranging from simple spelling errors to

Box 1. Some arbitrarily chosen examples that demonstrate various kinds of pitfalls in database usage

Synonyms

In organisms that are the target of major genetic studies, it often happens that the same gene is isolated by many different groups and so it ends up with many different names. For example, yeast *TUP1* is also known as *AER2, SPL2, CYC9, UME7, AAR1, AAM1* and *FLR1*. In *Escherichia coli*, *bnr* is also known as *bna4, dnaX, cmsZ, hgt1, may4, cur, pRG* and *hgt5*. The multiplicity of synonyms also exists at the level of protein names. For example, annexin V was also called: lipocortin V, endonexin II, calphobindin I, placental anticoagulant protein I, pp4, thromboplastin inhibitor, vascular anticoagulant-alpha and anchorin CII.

Different gene – same name

Conversely, it often happens that the same gene name is given to two different genes. Generally one of these duplicate names is quickly changed, but in some cases the two gene names each find a lobby and are simultaneously promoted. For example, yeast *MDP1* is both the gene for the mitochondrial peptide chain release factor 1 and for the mitochondrial respiratory function protein 1. A famous example is 'cyclin', the accepted name for a large family of cell-cycle components, which became so *prominent that this name* is no longer used for a protein now known as proliferating cell nuclear antigen (originally called cyclin).

Spelling

Even spelling mistakes can end up as gene synonyms. For example, the yeast gene, *SCG25* (suppressor of *CDC25*), was so often misquoted as *SDG25* that it has become an accepted synonym. In addition to spelling mistakes, database queries can be hindered by: differences between US and UK spelling (e.g. hemoglobin or haemoglobin); representation of special characters, such as accented characters (e.g. *Krüppel*, *Krueppel* or *Kruppel*); upper and lower cases (e.g. in the *Drosophila* genetic nomenclature, *H* is *Hairless*, but *h* is *hairy*).

Biological source and contamination

There are numerous problems with the annotation of the biological source of a sequence. For example, the OBGenelink division of EMBL/GenBank division should only contain sequences that are encoded on the mitochondria or plastid. But often entries reporting nuclear-encoded genes for proteins targeted to such an organelle are wrongly entered in the ORG division. The converse is also true as some chloroplast or mitochondrial encoded sequences are sometimes found in other divisions of EMBL/GenBank. This problem can have an effect on the derived protein sequence: if a nuclear-encoded mitochondrial gene is misclassified into the ORG section, the resultant translation will be wrong as the automatic translation software will assume that a mitochondrial genetic code should be used. The contamination of cDNA libraries (usually by fungal or bacterial DNA) is still an issue (a prominent example is one of Genethon's 'human' EST libraries that have a surprising number of matches with the yeast genome). Some scientific surprises can result from these issues: it was found recently (Laurent Duret, pers. commun.) that the sequence of two genes coding for annexin 1 and insulin from a sponge (their 'identification' as lower e[illegible]s was unexpected!) were too closely related to their mammalian homologs. It turned out that the biological samples had most probably been contaminated by an undetermined rodent species.

0168-9525/96/$15.00
PII: S0168-9525(96)80040-7

discrepancies between sequences published in printed form. This should decrease with electronic submissions and quality check software. Sequences can also be incorrectly labelled because of contamination of the sequenced material (Box 1).

A major annotation problem is the functional description of a gene. Due to the increasing number of 'software robots'[6] that automatically assign functions based on similarities, a single wrongly annotated entry will lead to whole families with artificial functions based on similarities to that entry [e.g. Ref. 7 in which information was carefully transferred from a gene called *nifr3* to several unannotated proteins; whether *nifr3* functions, however, in nitrogen fixation (nif) remains unclear, despite its annotation]. Automatic methods often cannot adapt to the dynamic annotation of entries and might, for example, point to neighboring genes in automatically translated protein sequences after a new ORF has been discovered in an already annotated region. Further annotation problems arise due to user interpretation (see below).

The retrieval of data is often hindered by incorrect genetic nomenclature (Box 1). Even in relatively similar organisms, such as budding yeast and fission yeast, the same gene name actually points to non-homologous functionally different proteins (e.g. *RAD4*). Details in the syntax can often not be reflected in the databases. Major attempts have been made for classifying enzymes, but even here the functional classification needs to be complemented by consistent homology and three-dimensional information. Nomenclature is also needed at the level of domains, structurally independent parts of proteins for which frequently several names coexist in the literature and consequently in the databases[8]. The different communities should force standardization.

Finally, databases are not always up to date regarding the functional information or other annotated features because there is currently no systematic update mechanism. Due to the policy of some databases that only authors can change the content of an entry, followup characterizations of genes or gene products are, thus, only occasionally included.

Problems of interpretation

Numerous pitfalls are related to the interpretation of the results of the database-accessing software; simple problems arise if the retrieval system does not access the full dataset so that stored information is not found. Furthermore, the occasional user often only accesses the major sequence databases that contain information from all organisms. Many communities studying particular protein families or organisms know about specialized databases that contain much more information on particular genes or proteins, but which is often not linked to the major databases. There is hope that this will change in the near future (e.g. links to FlyBase, YPD in SWISS-PROT).

A battery of traps result from database similarity searches, probably the most prominent form of database access. For example, the user might have insufficient knowledge of the limits of the programs (e.g. 'homology' to the coiled-coil region of myosin that is due to similar structural constraints) and inadequate thresholds and parameters often prevent an objective analysis. A different problem results from the pressure on sequencing groups, not to overlook interesting functional information in 'their' sequences. Thus, similarity search methods are stretched and spurious hits are taken as real. Moreover, similarities might only be restricted to certain domains, but the function is transferred to a whole protein. All such questionable interpretations end up in databases and are then considered as facts.

Finally, here is just one example that demonstrates the difficulties of functional predictions based on homology. Imagine the best hit to your *Drosophila* sequence is the human zinc-containing alcohol dehydrogenase class 4 μ/σ (in databases mu/sigma) chain (ADH7). It is very difficult to find out, whether the *Drosophila* sequence is the ortholog, another alcohol dehydrogenase, a homologous lactate dehydrogenase, a more distantly related oxidoreductase, or perhaps just a protein with an NADH-binding site. There is currently little quantification possible, in terms of functional similarity; a way out might be the knowledge of the complete

Box 2. Unusual database entries

'Protein' sequences in databases can be as short as one amino acid that is sometimes an X (as happens in the patent divisions of the databases) so that several database accession software packages have problems. Automatic DNA translation programs that contribute a considerable fraction of the protein sequences can also be mislead: the following TREMBL (automatic Translation of EMBL; version August 96) entry is a mistranslation (compare with the annotated CDS), probably due to some annotation problems in the corresponding EMBL entry. The name is also unusual: a human protein with an identifier starting with MM (usually meaning *Mus musculus*). It is supposed to encode a small region of trk4 but the translation comes up with parts of a different protein, MAC25. Detective work is needed to figure out the errors that lead to the wrong translation.

```
ID   MMTRK4A_1  standard; PRT; 118 AA.
AC   M55337;
DR   EMBL; M55337; MMTRK4A.
DE   gene: "trk4"; product: "oncogene tyrosine protein kinase receptor";
DE   Human oncogene tyrosine protein kinase receptor (trk4) mRNA,
DE   partial cds.
OS   Homo sapiens (human)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
FT   CDS             <1..>354
FT                   /gene="trk4"
FT                   /note="NCBI gi: 339916"
FT                   /codon_start=1
FT                   /product="oncogene tyrosine protein kinase receptor"
FT                   /db_xref="PID:g339916"
FT                   /translation="HSIKDNAALQALAQRQEEEKQKQEEQHQEEAATPSGGGRNRSAS
FT                   SSWVGTMAGISNSLNPNTLGGSSLSPPEGASGLQGHIIEEPQTFSDACVHKIQRDIV
FT                   LKWELGEGAFGKVF"
CC   translated using genetic code table "Standard"
CC   Warning: codon start shifted by 1
CC   Warning: illegal start codon
SQ   Sequence   118 BP;

Mmtrk4a_1  Length: 118  August 20, 1996 23:39  Type: P  Check: 7282  ..
       1  PLPPMPAMER PSLRALLLGA AGLLLLLLPL SSSSSSDTCG PCEPASCPPL
      51  PPLGCLLGET RDACGCCPMC ARGEGEPCGG GGAGRGYCAP GMECVKSRKR
     101  RKGKAGAAAG GPGVSGVC
```

GENETWORK

gene pool of organisms from all major taxa, which will allow classification within multigene families via phylogenetic trees.

Shared responsibilities

Although the list of problematic issues is much longer, we wish to point out that sequence databases are the most useful tool in sequence analysis and the question should be how can one further improve their value by enhancing the data storage, handling and retrieval? How should the responsibility for this task be shared? Everybody who stores information should feel responsible for the data and the annotation quality. Database teams have a restricted budget and can only provide some quality checks (e.g. for cloning and sequencing errors, artifically translated vectors, repeats and so on). Databases rely on standards and these have also to come from the different communities in the form of agreed nomenclature and clearly reproducible functional characterizations. Specialists should spend the time to give feedback on encountered problems and database teams should have mechanisms to include such improvements. This is, of course, easily said, but opinions about data and annotation vary and the truth is not always obvious. In conclusion, a concerted effort is needed from the database teams that have to maximize their services and the user community that should share responsibilities in taking care of the quality of the entries.

Peer Bork
bork@embl-heidelberg.de

EMBL, Meyerhofstr. 1, 69012 Heidelberg and Max-Delbrück-Center for Molecular Medicine, Berlin Buch, Germany.

Amos Bairoch
bairoch@cmu.unige.ch

Medical Biochemistry Department, University of Geneva, 1211 Geneva 4, Switzerland.

References

1 Bates, D.L (1992) *Trends Genet* 8, 52–55
2 Birney, E., Thompson, J. and Gibson, T. *Nucleic Acids Res.* (in press)
3 Bork, P. (1996) *Science* 271, 1431–1432
4 Hodi, R.W.W. *et al.* (1996) *Nature* 381, 272
5 Burset, M. and Guigo, R. (1996) *Genomics* 34, 353–367
6 Boguski, M. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71
7 Casari, G. *et al.* (1995) *Nature* 376, 647–648
8 Bork, P. and Bairoch, A. (1996) *Trends Biochem. Sci.* 20 (Suppl. C23) Perter

Genetwork is a regular column of news and information about Internet resources for researchers in genetics and development. Gen*Net is compiled and edited with the help of Steven E. Brenner (MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK. CB2 2QH) and Jeremy Rashbass (Department of Histopathology, Addenbrooke's Hospital, Hills Road, Cambridge, UK. CB2 2QQ).

If you would like to announce or publicize an Internet resource, please contact TIG@elsevier.co.uk

MEETING REPORTS

Flies in Crete

10TH EMBO WORKSHOP ON THE MOLECULAR AND DEVELOPMENTAL BIOLOGY OF *DROSOPHILA*, KOLYMARI, CRETE, 14–20 JULY 1996.

The results presented at this meeting were enormously varied and informative and the comments below represent just a small sample of the interesting science that was presented.

The importance of polarity within a single cell was illustrated several times. Transcripts of the segment polarity gene *inscuteable* (*insc*) were found to be apically localized in polarized epithelial cells (H. Kruse, Toronto) and, interestingly, localization was required for *insc* function. In contrast, localized *insc* transcripts were not required for *insc* function in neuroblast mesenchymal cells. Also, with regard to cellular polarity, the *inscuteable* gene was shown to be required for the orientation of the mitotic spindle and, therefore, for the correct plane of cell division (W. Chia, Singapore). Remarkably, in neuroblasts INSCUTEABLE is apically localized and is required for the basal localization of the homeodomain protein, PROSPERO. Thus, INSCUTEABLE appears to be an important component of the positional information within a cell.

Five major signaling pathways were discussed: *hedgehog* (*hh*), *wingless* (*wg*), *decapentaplegic* (*dpp*), EGF and FGF. Perhaps not surprisingly, several interactions and similarities between signaling pathways were apparent. For example, SMOOTHENED protein, which appears to be a G-coupled seven-transmembrane receptor, was suggested to be an HH receptor (M. Noll, Zürich). Curiously, SMOOTHENED shares striking similarity to the FRIZZLED family of proteins, which are putative receptors for Wnt (e.g. WG) signals. Downstream of the WG signal might be HMG-domain transcription factors related to mouse lymphoid enhancer-binding factor 1 (LEF1) (M. Bienz, Cambridge, UK, in collaboration with R. Grosschedl). LEF1 binds to ARMADILLO, another downstream component of the WG signal and, when expressed in *Drosophila*, phenocopies *wg*-overexpression phenotypes. Therefore, the tantalizing hypothesis that LEF1 might be an ARMADILLO-activated nuclear target for WG signaling was suggested. Another molecule downstream of the Wg signal is encoded by *arrow*. Surprisingly, *arrow* turns out to be identical to the gene *centrosomin* which is a component of the centrosome (T. Kaufman, Bloomington and S. DiNardo, New York). How the product(s) of a single gene can participate in two seemingly very different cellular processes provides a curious puzzle. In the embryonic endoderm, the homeodomain protein encoded by *extradenticle*, which binds to DNA cooperatively with HOX proteins, was shown to translocate from cytoplasm to nucleus in response to both DPP and WG signals (R. Mann, New York). Similarly, the novel protein encoded by *Mothers against dpp* was also shown to translocate from the cytoplasm into nuclei in response to DPP (W. Gelbart, Cambridge, USA). Thus, controlling protein localization within a cell might be a common response to extracellular signals.

In addition to intersecting signaling pathways, combinations of different signals were shown to be important for the activation of *even-skipped*-expression in a small cluster of mesoderm cells (A. Michelson, Cambridge, USA). The selection of these mesoderm cells, which are founders for a subset of muscles, depends on intersecting fields of *wg* and *dpp* expressing cells in the ectoderm, together with a RAS-dependent pathway (perhaps the EGF pathway). These signals define an equivalence group, which, as is the case for neuroblast selection, is refined by the action of another set of signaling molecules encoded by the neurogenic genes, *Notch* and *Delta*.

Evidence Appendix I

Alignment of LKR domains which is an alignment of the LKR domains of the plant bifunctional LKR/SDH proteins from Arabidopsis (SEQ ID NO:112), corn (SEQ ID NO:122, encoded by SEQ ID NO:120) and soybean (SEQ ID NO:121) and the monofunctional lysine-forming SDH proteins from *S.cerevisiae* (gi:453184), *C.albicans* (gi:1170847) and *Y.lipolytica* (gi:173262).

This was submitted as Appendix B which accompanied the Response submitted on July 20, 2007 and entered by Examiner in Office Communication dated March 5, 2008.

Appendix B

Appendix B shows a comparison of the amino acid sequences of the LKR domains of the bifunctional LKR/SDH proteins from Arabidopsis, corn and soybean, SEQ ID NOs: 112, 122 and 121, respectively, and the yeast monofunctional lysine-forming SDHs (gi: 453184, 1170847, and 173262). Amino acids conserved among at least one plant sequence and at least one yeast sequence are indicated with an asterix (*) on the top row; amino acids conserved among at least two plant sequences are indicated by a plus (+) above the alignment; dashes are used by the program to maximize the alignment of the sequences.

```
                          *               *   *        * *  ****     **  *          * *  *
                                                                  +++++ ++          ++ ++++
LKR domain of SEQ ID NO:112   MNSNGHEEEKKLGNGVVGILSETVNKWERRTPLTPSHCARLLHGG-KDRTGISRIVVQPS
LKR domain of SEQ ID NO:122   ------------------------------------CARLLLGGGKNGPRVNRIIVQPS
LKR domain of SEQ ID NO:121   ------------------------------------------------------------
GI: 453184                    MA----------AVTLHLRAETKPLEARAALTPTTVKKLIAKG-------FKIYVEDS
GI: 1170847                   MSKS----------PVILHLRAETKPLEARAALTPSTTKQLLDAG-------FEIYVEES
GI: 173262                    MT-A----------PVKLHLRAETKPLEHRSALTPTTTRKLLDAG-------FEVFVEKS

                                    *****  * *         *   *** *  ** *          * *  * *
                                  ++++++ ++  +++++ +     ++ ++++ ++++ + + +++  ++++++++++++
LKR domain of SEQ ID NO:112   AKRIHHDALYEHVGCEISDD-----LSDCGLILGIKQ-PELEMILPERAYAFFSHTHKAQ
LKR domain of SEQ ID NO:122   TRRIHHDAQYEDAGCEISED-----LSECGLIIGIKQ-PKLQMILSDRAYAFFSHTHKAQ
LKR domain of SEQ ID NO:121   ------------------------------------------------------------
GI: 453184                    PQSTFNINEYRQAGAIIVPAGSWKTAPRDRIIIGLKEMPETDTFPLVHEHIQFAHCYKDQ
GI: 1170847                   SQSTFDIKEYEAVGAKIVPEGSWKTAPKERIIFGLKELPENETFPLIHEHIQFAHCYKDQ
GI: 173262                    PLRIFDDQEFVDVGATLVEEGSWVSAPEDRMIIGLKELPE-ESFPLSHEHIQFAHCYKDQ

                                      *         ** * *    * * *  *** ***
                              ++++++++++ +++ + ++++++++ +++ +++++ ++++++ ++++++++++++++++
LKR domain of SEQ ID NO:112   KENMPLLDKILSERVTLCDYELIVGDHGKRLLAFGKYAGRAGLVDFLHGLGQRYLSLGYS
LKR domain of SEQ ID NO:122   KENMPLLDKILEERVSLFDYELIVGDDGKRSLAFGKFAGRAGLIDFLHGLGQRYLSLGYS
LKR domain of SEQ ID NO:121   ------------------------------------------------------------
GI: 453184                    AGWQNVLMRFIKGHGTLYDLEFLENDQGRRVAAFGFYAG----------------------FA
GI: 1170847                   AGWQDVLKRFPQGNGILYDLEFLENDQGRRVAAFGFYAG----------------------FA
```

```
GI: 173262                    GGWKDVLSRFPAGNGTLYDLEFLEDDNGRRVAAFGFHAG-------------------FA

                                            **      ** *  *                        * *
                              +++++++ + ++ ++++++++++ + ++++  +++ ++++ ++++++ ++++ +++++++
LKR domain of SEQ ID NO:112   TPFLSLGASYMYSSLAAAKAAVISVGEEIASQGLPLGICPLVFVFTGTGNVSLGAQEIFK
LKR domain of SEQ ID NO:122   TPFLSLGQSHMYPSLAAAKAAVIVVAEEIATFGLPSGICPIVFVFTGVGNVSQGAQEIFK
LKR domain of SEQ ID NO:121   ------------------------------------------------------------
GI: 453184                    GAALGV-RDWAFKQ---------THSDD---EDLP-AVSP----YPNEKAL---VKDVTK
GI: 1170847                   GAAIGV-LDWSFKQLNGNTKGTKGEGEG---GELP-GVTP----YPNENEL---IKDVKI
GI: 173262                    GAAIGV-ETWAFQQ---------THPDS---ENLP-GVSA----YPNETEL---VDKIKK


                               *        ****   *    *   *         *          *    *  *  * **
                              +++++++    ++++ +      +    +++++ + +++  ++ + +  ++    +++++
LKR domain of SEQ ID NO:112   LLPHTFV-EPSKLPELFVKDKGISQNGISTKRVYQVYGCIITSQDMVEHKDPSKSFDKAD
LKR domain of SEQ ID NO:122   LLPHTFV-DAEKLPEIF-QARNLSKQSQSTKRVFQLYGCVVTSRDIVSHKDPTRQFDKGD
LKR domain of SEQ ID NO:121   ------------------------------------------------EPKDHVIVFDKAD
GI: 453184                    DYKEALAT-GARKPTVLI----IGALGRCGSGAIDLL------HKVGIPDANILKWDIKE
GI: 1170847                   ELEKALTKNGGQYPKCLV----IGALGRCGSGAIDLF------KKIGIPDDNIAKWDMAE
GI: 173262                    DLAAAVEK-GSKLPTVLV----IGALGRCGSGAIDLA------RKVGIPEENIIRWDMNE


                                        * *     *    ** *  *  **  *** *     *    * *   *
                              ++++++++++++++++++++++++++++++++++ +++ +++++++  + ++++++++++
LKR domain of SEQ ID NO:112   YYAHPEHYNPVFHEKISPYTSVLVNCMYWEKRFPCLLSTKQLQDLTKKGLPLVGICDITC
LKR domain of SEQ ID NO:122   YYAHPEHYTPVFHERIAPYASVIVNCMYWEKRFPPLLNMDQLQQLMETGCPLVGVCDITC
LKR domain of SEQ ID NO:121   YYSHPEHYNPTFHEKIAPYASVIVNCMYWEKRFPQLPSYKQMQDLMGRGSPLVGIADITC
GI: 453184                    -----TSRGGPFDEIPQ--ADIFINCIYLSKPIAPFTNMEKLNNPNRRLRTVVDVSADTT
GI: 1170847                   -----TAKGGPFQEIVD--SDIFINCIYLSKPIPPFINKEILNNENRKLTTIVDVSADTT
GI: 173262                    -----TKKGGPFQEIAD--ADIFINCIYLSQPIPPFINYDLLNKETRKLSVIVDVSADTT

                                   *     **          **          *       *    * **      ***  *
                              ++++++++++ ++++++++++++ +++++++++ ++ ++++++++++++++++++++
LKR domain of SEQ ID NO:112   DIGGSIEFVNRATLIDSPFFRFNPSNNSYYDDMDGDGVLCMAVDILPTEFAKEASQHFGD
LKR domain of SEQ ID NO:122   DIGGSIEFINKSTSIERPFFRYDPSKNSYHDDMEGAGVVCLAVDILPTEFSKEASQHFGN
LKR domain of SEQ ID NO:121   DIGGSIEFVNRGTSIDSPFFRYDPLTNSYHDDMEGNGVICLAVDILPTEFAKEASQHFGN
GI: 453184                    NPHNPIPIYTVATV-------FNKPTVLVPTTVGPKLSVI-SIDHLPSLLPREASEFFSH
GI: 1170847                   NPHNPIPVYEIATV-------FNEPTVEVKLDKGPKLSVC-SIDHLPSLLPREASEFFAK
GI: 173262                    NPHNPVPVYTIATT-------FDHPTVPVETTAGPKLSVC-SIDHLPSLLPREASEAFSE
```

```
                                  *     **    *    *   *              * *    *  **
                                 +++  +  +++      ++++++++++++ + ++++++ ++++++ +
LKR domain of SEQ ID NO:112      ILSGFVGSLASMTEISDLPAHLKRACISYRGELTSLYE-YIPRMRKSNPE
LKR domain of SEQ ID NO:122      ILSRLVASLASVKQPAELPSYLRRACIAHAGRLTPLYE-YIPRMRNTMID
LKR domain of SEQ ID NO:121      ILSQFVVNLASATDITKLPAHLRRACIAHKGVLTSLYD-YIPRMRSSDS-
GI: 453184                       DL---LPSLELLPQRKTAPVWVRA---------KKLFDRHCARVKRSSRL
GI: 1170847                      DL---MPSLLELPNRDTSPVWVRA---------KQLFDKHVARLDKE---
GI: 173262                       AL---LPSLLQLPQRDTAPVWTRA---------KALFDKHVLRIGE----
```

Evidence Appendix J

Comparison of the SDH domains of the bifunctional plant LKR/SDH proteins from Arabidopsis (SEQ ID NO:112), corn (SEQ ID NO:122, encoded by SEQ ID NO:120) and soybean (SEQ ID NO:121) and the monofunctional glutamate-forming SDH protein from *S.cerevisiae* (gi:729968).

This was submitted as Appendix C which accompanied the Response submitted on July 20, 2007 and entered by Examiner in Office Communication dated March 5, 2008.

Appendix C

Appendix C shows a comparison of the amino acid sequences of the SDH domains of the bifunctional LKR/SDH proteins from Arabidopsis, corn and soybean, SEQ ID NOs: 112, 122 and 121, respectively, and the yeast monofunctional glutamate-forming SDH (gi:729968). Amino acids conserved among at least one plant sequence and the yeast sequence are indicated with an asterix (*) on the top row; amino acids conserved among at least two plant sequences are indicated by a plus (+) above the alignment; dashes are used by the program to maximize the alignment of the sequences.

```
                                *   ***** * *  *  * **                        *** ****   *
                                ++  ++++++++++++++++++++  +++ +++++    + +  ++++++++++++++
SDH domain of SEQ ID NO:112     MTKKSGVLILGAGRVCRPAADFLASVRTISSQQWYKTYFGADSEEKTDVHVIVASLYLKD
SDH domain of SEQ ID NO:122     GPK---ILILGAGRVCRPAAEFLASYPDICT-------YGVDDHDADQIHVIVASLYQKD
SDH domain of SEQ ID NO:121     ---KAAVLILGAGRVCQPAAEMLSSFGRPSSSQWYKTLLEDDFECQTDVEVIVGSLYLKD
GI: 729968                      MGKN--VLLLGSGFVAQPVIDTLAA----------------------NDDINVTVACRTLAN

                                * *     *   ** *** *   * *     ********  **  * *  *  *   **
                                ++++++++ +++++++++ +  ++ ++++++++++++++++++ ++  ++++++++++
SDH domain of SEQ ID NO:112     AKETVEGISDVEAVRLDVSDSESLLKYVSQVDVVLSLLPASCHAVVAKTCIELKKHLVTA
SDH domain of SEQ ID NO:122     AEETVDGIENTTATQLDVADIGSLSDLVSQVEVVISLLPASFHAAIAGVCIELKKHMVTA
SDH domain of SEQ ID NO:121     AEQTVEGIPNVTGIQLDVMDRANLCKYISQVDVVISLLPPSCHIIVANACIELKKHLVTA
GI: 729968                      A-QALAKPSGSKAISLDVTDDSALDKVLADNDVVISLIPYTFHPNVVKSAIRTKTDVVTS

                                **      *      ****   * ********* * * *** *   ** **********
                                ++++ +++++  ++++++++++++++++++++++++++++++ ++  ++++++++++++++
SDH domain of SEQ ID NO:112     SYVDDETSMLHEKAKSAGITILGEMGLDPGIDHMMAMKMINDAHIKKGKVKSFTSYCGGL
SDH domain of SEQ ID NO:122     SYVDESMSNLSQAAKDAGVTILCEMGLDPGIDHLMSMKMIDEAHARKGKIKAFTSYCGGL
SDH domain of SEQ ID NO:121     SYVDSSMSMLNDKAKDAGITILGEMGLDPGIGHMMAMKMINQAHVRKGKIKSFTSYCGGL
GI: 729968                      SYISPALRELEPEIVKAGITVMNEIGLDPGIDHLYAVKTIDEVHRAGGKLKSFLSYCGGL

                                * **   *** *****   * * * ** ***   * *  *   ** * **       *
                                ++++++++++++++++++++++++ +++ ++  ++++++++ ++++++ + + + ++++
SDH domain of SEQ ID NO:112     PSPAAANNPLAYKFSWNPAGAIRAGQNPAKYKSNGDIIHVDGKNLYDSAARFRVPNLPAF
SDH domain of SEQ ID NO:122     PSPAAANNPLAYKFSWNPAGALRSGKNPAVYKFLGETIHVDGHNLYESAKRLRLRELPAF
SDH domain of SEQ ID NO:121     PSPEAANNPLAYKFSWNPAGAIRAGRNPATYKWGGETVHIDGDDLYDSATRLRLPDLPAF
GI: 729968                      PAPEDSDNPLGYKFSWSSRGVLLALRNSAKYWKDGKIETVSSEDLMATAKPYFI--YPGY
```

```
                                *  * *****  *  *** *  ** *  ******  ** *    *   *     **
                                ++++++++++ ++++++++ ++++++++++++++++++++ ++++ ++++ +++ ++  +
SDH domain of SEQ ID NO:112     ALECFPNRDSLVYGEHYGIESEATTIFRGTLRYEGFSMIMATLSKLGFFDSEANQVLSTG
SDH domain of SEQ ID NO:122     ALEHLPNRNSLIYGDLYGISKEASTIYRATXRYEGFSEIMVTLSKTGFFDAANHPLLQDT
SDH domain of SEQ ID NO:121     ALECLPNRNSLLYGDLYGI-TEASTIFRGTLRYEGFSEIMGTLSRISLFNNEAHSLLMNG
GI: 729968                      AFVCYPNRDSTLFKDLYHIP-EAETVIRGTLRYQGFPEFVKALVDMGMLKDDANEIF--S

                                * *     *   *             *      ***** *      ** *
                                 +++++ ++ ++++    +  +++ +    + ++  +++++++++   ++ + ++++ +
SDH domain of SEQ ID NO:112     KRITFGALLSNILNKDADNESEPLAG-----EEEISKRIIKLGHSKE--TAAKAAKTIVF
SDH domain of SEQ ID NO:122     SRPTYKGFLDELLNNISTINTDLDIEASGGYDDDLIARLLKLGCCKNKEIAVKTVKTIKF
SDH domain of SEQ ID NO:121     QRPTFKKFLFELLKVVGDNPDELLIG-----ENDIMEQILIQGHCKDQRTAMETAKTIIF
GI: 729968                      KPIAWNEALKQYL------------GAKSTSKEDLIASIDSKATWKDDEDRERILSGFAW

                                ***     *      * *  * * ** * *  ** *** * *    *  **  **  * *
                                +++ ++++++  +++ +++ + ++++++++  ++++++++++++++++++++  +++ ++
SDH domain of SEQ ID NO:112     LGFNEEREVPSLCKSVFDATCYLMEEKLAYSGNEQDMVLLHHEVEVEFLESKRIEKHTAT
SDH domain of SEQ ID NO:122     LGLHEETQIPKGCSSPFDVICQRMEQRMAYGHNEQDMVLLHHEVEVEYPDGQPAEKHQAT
SDH domain of SEQ ID NO:121     LGLLDQTEIPASCKSAFDVACFRMEERLSYTSTEKDMVLLHHEVEIEYPDSQITEKHRAT
GI: 729968                      LGLFSDAKITPR-GNALDTLCARLEELMQYEDNERDMVVLQHKFGIEWADGT-TETRTST

                                *   ***         ** *** * **     *   **  * * *  ***  *   *
                                ++++++  +  ++++++++++++++++++++++++++++++++++++++++++++ +++ +++++
SDH domain of SEQ ID NO:112     LLEFGDIKNGQTTTAMAKTVGIPAAIGALVLIEDKIKTRGVLRPLEAEVYLPAL-DILQA
SDH domain of SEQ ID NO:122     LLEFGKVENGRSTTAMALTVGIPAAIGALLLLKNKVQTKGVIRPLQPEIYVPAL-EILES
SDH domain of SEQ ID NO:121     LLEFGKTLDEKTTTAMALTVGIPAAVGALLLLTNKIQTRGVLRPIEPEVYNPAL-DIIEA
GI: 729968                      LVDYGKVG---GYSSMAATVGYPVAIATKFVLDGTIKGPGLLAPYSPEINDPIMKELKDK

                                *** * ***
                                +++++ ++ +
SDH domain of SEQ ID NO:112     YGIKLMEKAE.
SDH domain of SEQ ID NO:122     SGIKLVEKVET
SDH domain of SEQ ID NO:121     YGIKLIEKT-E
GI: 729968                      YGIYLKEKTVA
```

Evidence Appendix K

This is an alignment of the plant bifunctional proteins from Arabidopsis, corn and soybean, SEQ ID NOs:112, 122 and 121, respectively.

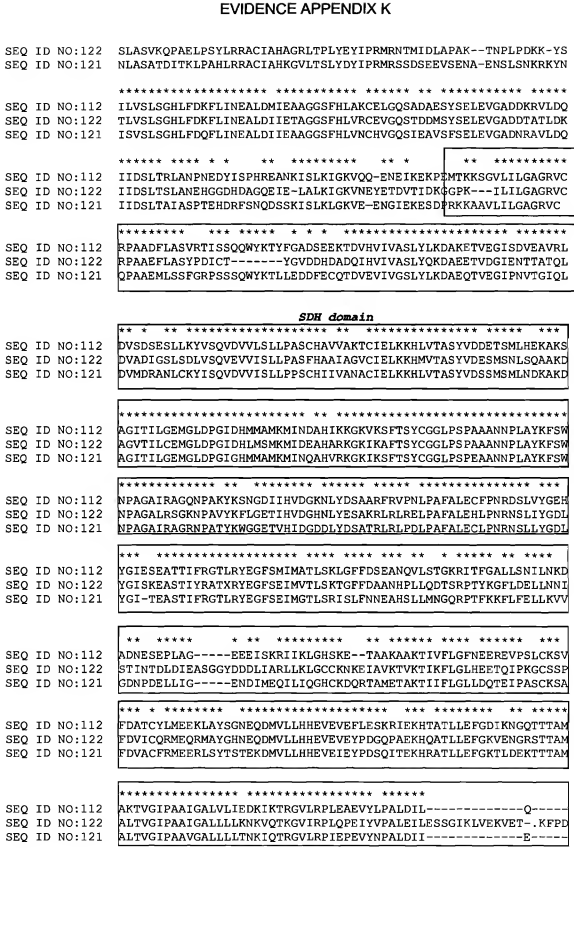
This was submitted as Appendix D which accompanied the Response submitted on July 20, 2007 and entered by Examiner in the Office Communication dated March 5, 2008.

EVIDENCE APPENDIX K

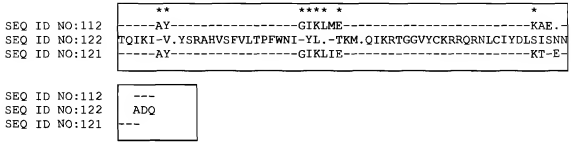
Appendix D

Appendix D shows a comparison of the amino acid sequences of the bifunctional LKR-SDH proteins from Arabidopsis, corn and soybean, SEQ ID NOs: 112, 122 and 121, respectively. Amino acids conserved among at least two plant sequences are indicated with an asterix (*) on the top row; dashes are used by the program to maximize the alignment of the sequences. The LKR and SDH domains (boxed sequences) were identified by Epelbaum et al. (Plant Mol. Biol. 35:735-748 (1997)) and Tang et al. (Plant Cell 9:1305-1316 (1997))

```
                                                  ***** ** *      ** ****
SEQ ID NO:112  MNSNGHEEEKKLGNGVVGILSETVNKWERRTPLTPSHCARLLHGG-KDRTGISRIVVQPS
SEQ ID NO:122  ------------------------------------CARLLLGGGKNGPRVNRIIVQPS
SEQ ID NO:121  ------------------------------------------------------------

                 ****** **  ***** *** ********** * ***  ******************
SEQ ID NO:112  AKRIHHDALYEHVGCEISDDLSDCGLILGIKQPELEMILPERAYAFFSHTHKAQKENMPL
SEQ ID NO:122  TRRIHHDAQYEDAGCEISEDLSECGLIIGIKQPKLQMILSDRAYAFFSHTHKAQKENMPL
SEQ ID NO:121  ------------------------------------------------------------

               ***** *** * ******** *** ***** ****** ********************
SEQ ID NO:112  LDKILSERVTLCDYELIVGDHGKRLLAFGKYAGRAGLVDFLHGLGQRYLSLGYSTPFLSL
SEQ ID NO:122  LDKILEERVSLFDYELIVGDDGKRSLAFGKFAGRAGLIDFLHGLGQRYLSLGYSTPFLSL
SEQ ID NO:121  ------------------------------------------------------------

                                 LKR domain
               * * ** ********** * ****  *** **** ****** **** *************
SEQ ID NO:112  GASYMYSSLAAAKAAVISVGEEIASQGLPLGICPLVFVFTGTGNVSLGAQEIFKLLPHTF
SEQ ID NO:122  GQSHMYPSLAAAKAAVIVVAEEIATFGLPSGICPIVFVFTGVGNVSQGAQEIFKLLPHTF
SEQ ID NO:121  ------------------------------------------------------------

               *   **** *       *    ***** * ***  ** * ******   ***********
SEQ ID NO:112  VEPSKLPELFVKDKGISQNGISTKRVYQVYGCIITSQDMVEHKDPSKSFDKADYYAHPEH
SEQ ID NO:122  VDAEKLPEIF-QARNLSKQSQSTKRVFQLYGCVVTSRDIVSHKDPTRQFDKGDYYAHPEH
SEQ ID NO:121  ------------------------------------EPKDHVIVFDKADYYSHPEH

               ************************** *** *******  * ****************
SEQ ID NO:112  YNPVFHEKISPYTSVLVNCMYWEKRFPCLLSTKQLQDLTKKGLPLVGICDITCDIGGSIE
SEQ ID NO:122  YTPVFHERIAPYASVIVNCMYWEKRFPPLLNMDQLQQLMETGCPLVGVCDITCDIGGSIE
SEQ ID NO:121  YNPTFHEKIAPYASVIVNCMYWEKRFPQLPSYKQMQDLMGRGSPLVGIADITCDIGGSIE

               **** ************* ********* ** ********************** **
SEQ ID NO:112  FVNRATLIDSPFFRFNPSNNSYYDDMDGDGVLCMAVDILPTEFAKEASQHFGDILSGFVG
SEQ ID NO:122  FINKSTSIERPFFRYDPSKNSYHDDMEGAGVVCLAVDILPTEFSKEASQHFGNILSRLVA
SEQ ID NO:121  FVNRGTSIDSPFFRYDPLTNSYHDDMEGNGVICLAVDILPTEFAKEASQHFGNILSQFVV

               **** * *  *********** * *********** *  *** *   * ** ** *
SEQ ID NO:112  SLASMTEISDLPAHLKRACISYRGELTSLYEYIPRMRKSNPEEAQDNIIANGVSSQRTFN
```


SEQ ID NO:122 SLASVKQPAELPSYLRRACIAHAGRLTPLYEYIPRMRNTMIDLAPAK--TNPLPDKK-YS
SEQ ID NO:121 NLASATDITKLPAHLRRACIAHKGVLTSLYDYIPRMRSSDSEEVSENA-ENSLSNKRKYN
SEQ ID NO:112 ILVSLSGHLFDKFLINEALDMIEAAGGSFHLAKCELGQSADAESYSELEVGADDKRVLDQ
SEQ ID NO:122 TLVSLSGHLFDKFLINEALDIIETAGGSFHLVRCEVGQSTDDMSYSELEVGADDTATLDK
SEQ ID NO:121 ISVSLSGHLFDQFLINEALDIIEAAGGSFHLVNCHVGQSIEAVSFSELEVGADNRAVLDQ
SEQ ID NO:112 IIDSLTRLANPNEDYISPHREANKISLKIGKVQQ-ENEIKEKPEMTKKSGVLILGAGRVC
SEQ ID NO:122 IIDSLTSLANEHGGDHDAGQEIE-LALKIGKVNEYETDVTIDKGGPK---ILILGAGRVC
SEQ ID NO:121 IIDSLTAIASPTEHDRFSNQDSSKISLKLGKVE-ENGIEKESDPRKKAAVLILGAGRVC
SEQ ID NO:112 RPAADFLASVRTISSQQWYKTYFGADSEEKTDVHVIVASLYLKDAKETVEGISDVEAVRL
SEQ ID NO:122 RPAAEFLASYPDICT-------YGVDDHDADQIHVIVASLYQKDAEETVDGIENTTATQL
SEQ ID NO:121 QPAAEMLSSFGRPSSSQWYKTLLEDDFECQTDVEVIVGSLYLKDAEQTVEGIPNVTGIQL
SDH domain
SEQ ID NO:112 DVSDSESLLKYVSQVDVVLSLLPASCHAVVAKTCIELKKHLVTASYVDDETSMLHEKAKS
SEQ ID NO:122 DVADIGSLSDLVSQVEVVISLLPASFHAAIAGVCIELKKHMVTASYVDESMSNLSQAAKD
SEQ ID NO:121 DVMDRANLCKYISQVDVVISLLPPSCHIIVANACIELKKHLVTASYVDSSMSMLNDKAKD
SEQ ID NO:112 AGITILGEMGLDPGIDHMMAMKMINDAHIKKGKVKSFTSYCGGLPSPAAANNPLAYKFSW
SEQ ID NO:122 AGVTILCEMGLDPGIDHLMSMKMIDEAHARKGKIKAFTSYCGGLPSPAAANNPLAYKFSW
SEQ ID NO:121 AGITILGEMGLDPGIGHMMAMKMINQAHVRKGKIKSFTSYCGGLPSPEAANNPLAYKFSW
SEQ ID NO:112 NPAGAIRAGQNPAKYKSNGDIIHVDGKNLYDSAARFRVPNLPAFALECFPNRDSLVYGEH
SEQ ID NO:122 NPAGALRSGKNPAVYKFLGETIHVDGHNLYESAKRLRLRELPAFALEHLPNRNSLIYGDL
SEQ ID NO:121 NPAGAIRAGRNPATYKWGGETVHIDGDDLYDSATRLRLPDLPAFALECLPNRNSLLYGDL
SEQ ID NO:112 YGIESEATTIFRGTLRYEGFSMIMATLSKLGFFDSEANQVLSTGKRITFGALLSNILNKD
SEQ ID NO:122 YGISKEASTIYRATXRYEGFSEIMVTLSKTGFFDAANHPLLQDTSRPTYKGFLDELLNNI
SEQ ID NO:121 YGI-TEASTIFRGTLRYEGFSEIMGTLSRISLFNNEAHSLLMNGQRPTFKKFLFELLKVV
SEQ ID NO:112 ADNESEPLAG-----EEEISKRIIKLGHSKE--TAAKAAKTIVFLGFNEEREVPSLCKSV
SEQ ID NO:122 STINTDLDIEASGGYDDDLIARLLKLGCCKNKEIAVKTVKTIKFLGLHEETQIPKGCSSP
SEQ ID NO:121 GDNPDELLIG-----ENDIMEQILIQGHCKDQRTAMETAKTIIFLGLLDQTEIPASCKSA
SEQ ID NO:112 FDATCYLMEEKLAYSGNEQDMVLLHHEVEVEFLESKRIEKHTATLLEFGDIKNGQTTTAM
SEQ ID NO:122 FDVICQRMEQRMAYGHNEQDMVLLHHEVEVEYPDGQPAEKHQATLLEFGKVENGRSTTAM
SEQ ID NO:121 FDVACFRMEERLSYTSTEKDMVLLHHEVEIEYPDSQITEKHRATLLEFGKTLDEKTTTAM
SEQ ID NO:112 AKTVGIPAAIGALVLIEDKIKTRGVLRPLEAEVYLPALDIL----------Q-----
SEQ ID NO:122 ALTVGIPAAIGALLLLKNKVQTKGVIRPLQPEIYVPALEILESSGIKLVEKVET-.KFPD
SEQ ID NO:121 ALTVGIPAAVGALLLLTNKIQTRGVLRPIEPEVYNPALDII----------E-----

EVIDENCE APPENDIX K



Related Proceedings Appendix

None